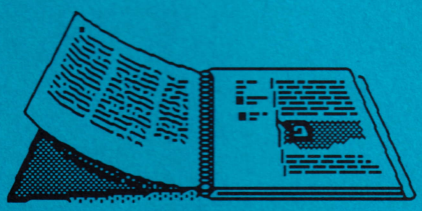
Zentrum für Allgemeine Sprachwissenschaft,
Sprachtypologie und Universalienforschung

ZAS Papers in Linguistics

Volume 11
September 1998



Edited by

Artemis Alexiadou
Nanna Fuhrhop
Ursula Kleinhenz
Paul Law

ISSN 1435-9588

The ZAS Papers in Linguistics was originally published by the Forschungsschwerpunkt Allgemeine Sprachwissenschaft, Typologie und Universalienforschung (FAS, Research Center for General Linguistics, Typology and Universals). The Center is now known as Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung (ZAS) under the auspices of the Deutsche Forschungsgemeinschaft (The German Research Foundation) and the Land of Berlin. The Center currently has research projects in syntax, semantics, morphology, phonology, phonetics as well as language change and language acquisition. ZAS provides a forum for the exchange of ideas in the academic community of the Berlin area through lectures, seminars, workshops and conferences. The Center cooperates with other universities in Germany, and sponsors visits by scholars from Europe and America.

Director: Ewald Lang

For further information on the Center, please write to:

Ewald Lang, Director
ZAS
Jägerstr. 10/11
D-10117 Berlin
Germany

Telephone: +49 30 20 19 24 01
Fax: +49 30 20 19 24 02

E-mail: sprach@fas.ag-berlin.mpg.de

ZAS Papers in Linguistics is intended to reflect the on-going work at the Center. It publishes papers by its staff members and visiting scholars. The publication is available on an exchange basis. For further information, please write to:

The editors
ZAS Papers in Linguistics
Jägerstr. 10/11
D-10117 Berlin
Germany

Telephone: +49 30 20 19 24 04
Fax: +49 30 20 19 24 02

E-mail: faspil@fas.ag-berlin.mpg.de

Papers of the Conference on 'The word as a phonetic unit'

ZAS, Berlin, October 22 - 23, 1997

Preface

This issue of ZASPIL contains the partially revised papers read at the conference on "The word as a phonetic unit". This workshop meeting was held on 22 - 23 October 1997 at the ZAS prior to the "Conference on the phonological word"[1], which took place over the following three days.

Whereas in phonology the (phonological) word constitutes a genuine domain the status of the word is all but clear from the phonetic point of view. This is also reflected in the quite diverse contributions to this volume. These were loosely grooped according to their main focus:

At the outset we put the two contrastive reports of H.G. Tillmann (Munich), who points to the central role of the word for natural as well as automatic speech recognition, and of K. Kohler (Kiel) on the effective disappearance of words in connected speech.

The second section is dedicated to the pronunciation variants observable in large speech corpora. P. Keating's (Los Angeles) paper deals with the coverage of the most common pronunciation and pronunciation variants in different speech corpora of American English. The contribution of F. Schiel and A. Kipp (Munich) provides an overview of an automatic segmentation procedure for German speech data based on rule generated pronunciation variants whereas Ch. Draxler (Munich) describes a PROLOG based data base management system for accessing segmental signal data, transcriptions, citation forms and orthographic representations of the words for German data.

The third session focusses on suprasegmental aspects: A. Mengel (Berlin) reports on his work on the location and realization of German lexical stress and R. Benzmüller & M. Grice (Saarbrucken) ask the question whether the low tone of the nuclear accentual fall in German is to be described as connected to the word or rather is dependent on the phrase structure of the utterance.

The fourth section on phonetic and/or phonological processes consists of A. Simpson's (Kiel) paper on the phonetics of German r, in which he argues in favour of an approach that doesn't imply phonological processes, the contribution of L. Ellis & W. Hardcastle (Edinburgh) on an EPG study of word final nasal alveolar to velar assimilation in English, and the paper of K. Russell (Manitoba) questioning the existence of Cree external sandhi as an overall categorical effect.

The fifth section focusses on articulatory word boundary phenomena. E. Farnetani (Padova) reports of combined EPG and acoustic studies showing different articulatory realisations of /t/ at word boundaries in Italian whereas in his paper on a transillumination study of voiceless obstruents in German M. Jessen (Stuttgart) questions the influence of word boundaries on the glottal devoicing behaviour.

The last session is dedicated to phonetic variation in German. In the first paper B. Kröger (Cologne) exemplifies observable word form variations that can be modelled by temporal variation within his gestural model. The contribution of L. Faust (Bonn) discusses the different phonetic variations in connection with speaking style. The final paper by B. Pompino-Marschall & P. Janker (Berlin) reports on first results of a study on the production and perception of the German word forms *'ein'* and *'einen'* and their different degrees of rduction.

At the end of this volume we added a list of the participants of this meeting with their affiliations as well as their postal and electronic addresses.

---

[1] Selected papers from this conference, organized by T.A. Hall and U. Kleinhenz, will be published as "Studies on the phonological word" (Amsterdam/Philadelphia: John Benjamins)

We once again want to thank the speakers/authors and all the participants of this conference for contributing to a stimulating discussion on the status of the word in phonetics and hope that in the near future we will have a new opportunity to meet here at the ZAS to exchange our ideas on central issues of phonetics and its 'linguistic neighbors'.

Last but not least we want to thank the ZAS for making this meeting possible and the Senate of Berlin and the German Research Council for funding.

Berlin, Summer 1998

Bernd Pompino-Marschall
Christine Mooshammer

Why the word should become the central unit of phonetic speech research

Hans G. Tillmann
IPSK, Universität München

The purpose of this paper is

(i) to assert that phonetics as speech science needs a fundamentally new theoretical orientation (leading also to a quite new research agenda);

(ii) to argue that this can be most easily[1] achieved if the word which to all native and non-native speakers of a language is the most naturally given unit of speech should also be considered the most central unit of speech science;

(iii) to promote the idea that phonetics - taken in this way and in close connection with the other basic speech science, semantics - will become a much more useful discipline in the context of the development of speech technologies for the coming so-called information society; and finally

(iv) to present along with earlier ideas published by the present author a rather complete picture of the phonetic processes which are basic in the production of natural speech acts.

My personal impression concerning the present situation of speech technology and spoken language processing is that despite all the remarkable successes of HMM- and NN-applications, a local minimum has been reached. For the further development of speech technology it will not be enough just to collect still more data for the training and testing of SLP-systems. Much more money has to be invested into new speech research and further development of theoretical concepts is needed. Therefore, and besides introducing the word as the central unit of phonetic speech research, I will also explicitely reintroduce the classical experimental phonetic concept of *systematic modification* (as apposed to random statistical variation of the phonetic forms of utterances[2]).

---

[1] In this paper the terms 'easy' and 'simple' are used with the meaning of Herrmann Paul's (1898:4) statement "dass ich nur für diejenigen schreibe, die mit mir der Überzeugung sind, dass die Wissenschaft nicht vorwärts gebracht wird durch komplizierte Hypothesen, mögen sie auch mit noch soviel Geist und Scharfsinn ausgeklügelt sein, sondern durch einfache Grundgedanken, die an sich evident sind, die aber erst fruchtbar werden, wenn sie zu klarem Bewußtsein gebracht und mit strenger Konsequenz durchgeführt werden".

[2] The discovery of certain prosodic sound variations (in words like *pâte, pâté, pâtisserie*) caused Rousselot, more than 100 years ago, to found the new discipline of Experimental Phonetics devoted to "les modifications phonétiques du langage".

The paper is organized in the following way. We first discuss the relationship between speaking and writing with respect to the word as a phonetic category. Then we try to clarify - in three sections dealing with considerations of priciple - the concept of the word as something that can be systematically modified in its phonetic form. In the second part we are going to give three more concrete examples of my basic idea (of systematically modifying the phonetic forms of the words of the given language). Finally we would like to come back to the traditional questions mainly asked by linguists theoretically concerning simplicity and complexity. Concern with simplicity and complexity of speech processes is one thing, another is to consider the interrelations between simplicity and complexity as a phonetic matter of fact when trying to understand the processes of phonetically producing an act of speech. Thus we have the following eight sections:

Introduction:

(1) Speaking and writing

Part I: Considerations of principle

(2) Sound words and language words

(3) Autonymic sound words and hetronymic language words in speech technology (SLP)

(4) Random variation and systematic modification

Part II: Three examples of practical applications

(5) Articulation of complex and elementary sound words

(6) CRIL and the central role of the word in the preparation of databases for SLP-technologies

(7) PHD: From single words to connected speech

Concluding remarks:

(8) On simplicity and complexity

As most of the color pictures and figures presented at the conference in Berlin can not be shown in the printed version, we have prepared a web-version which the reader may find on my home page at the following address:

http://www.phonetik.uni-muenchen.de/

1. Speaking and writing

As an introduction to what follows I first would like to develop an idea of what - in a narrower phonetic sense - could be called a word (of a given language). We start by looking at the complex relationship that exists between speaking and writing (both involving very complicated, mutually interrelated skilled human actions), and then create a point of departure for the following seven sections.

If we consider what speakers and writers of an utterance are actually producing as pure data, these are so different in nature and form, we would not be able to discover any specific proper-

ties in the written symbols and the recorded speech sounds which could be used to interrelate the two types of data to each other. It is only when we introduce the natural concept of the words of a given language and recognize that we need to know how the words of this language are to be pronounced and how they are to be written that we are in a position to compare any pair of spoken and written phrase of a language. If we recognize that a spoken and written utterance of this language contain exactly the same words in the same order are we inclined to say that they are identical - despite the fact that they are so different in nature and form.[3]

Two written or printed sequences of words such as 'I'm hungry' and 'I'm hungry' are identical if they contain the same sequence of letters including spaces. But it is also true that two phonetically very differently produced utterances of speech are said to be alike (or even identical) if they can be written as the same sequence of words (such as Bloomfield's "I'm hungry" uttered by a child who has eaten and merely wants to put off going to bed or uttered by a needy stranger at the door who wants to express 'please give me something to eat').

In any kind of speech science words are needed to explain the semantic relations that are created by the speaker who expresses himself in an act of speech. These semantic relations exist between the directly observable utterances and what the speaker wants to say (which is not directly observable). And these relations determine what the listener understands and infers when perceiving what the speaker is producing in a given context. This also seems to remain true in a rather technical situation.

When trained listeners look at spontaneously uttered speech data collected for the training of SLP-systems, the production of an orthographic transliteration becomes extremly difficult (if not nearly impossible) as soon as the listener cannot decide which word or sequence of words a given speaker has been trying to utter. In the VERBMOBIL-project, together with our partners in Kiel and Bonn, we have developed tools for handling cases where a word is mispronounced, mutilated, or even unrecognizable. On the other hand, in real speech acts, where we only want to understand the speaker (without having to produce a transliteration of what he is literally saying) we normally ignore these unclear parts of an utterance. Quite automatically we even infer mutilated words from a given context if these words are required for understanding what a speaker is saying. The only necessary condition for any natural speech act seems to be that (i) there is a speaker who creates certain phonetic facts (directly observed by the speaker himself and an audience, if present), and that (ii) the semantic facts (which, with the exception of so-called pairing situations, cannot normally be observed directly) are to be inferred in dependency of observable properties of the speakers utterance, and that (iii) these observable properties can be related to phonetically reproducible words of the language of the speaker.

To provide a point of departure for the following sections I would like to add a new technical detail to my earlier analysis concerning the semantics of phonetic transcription. This small, but helpful detail is simply to distinguish between the different use of two types of quotation marks.

In their analysis of human speech acts logicians and philosophers of language have introduced

---

[3] I have tried to specify the eight main differences between collections of spoken and written language data in Tillmann 1997.

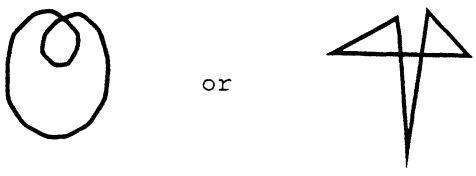
the concept of an ***autonymic*** use of some categorically established entities. If these entitities are written down orthographically quotation marks are employed for indicating an autonymic representation of an utterance. Thus they indicate that a speaker who utters "it is raining" tells the truth iff it is raining. This use of quotation marks as well as the word 'iff' (written with two f-letters to get a shorter version of 'if and only if') may indicate how much philosophers and logicians need writing for their analysis of speech acts! (Which, of course, is true for any kind of speech science; without the invention of writing systems only a few k years ago there would be no science in today's sense at all.)

In my own analysis of phonetic transcription as a way of symbolically representing directly observable phonetic events by graphically well defined entitities the starting idea[4] was to show that Gerold Ungeheuer's "extrakommunikative Situationen" may include very specific speech acts in which what the speaker wants to communicate by producing an audible utterance is nothing else but a demonstration of the phonetic form of this utterance. In such a situation the phonetic event is produced by the speaker in an autonymic way, meaning itself.

In speech we don't have the option to employ quotation marks to indicate such an autonymic use of a phonetic event as a reproducible category of its own. Therefore we have to find another solution.

It was here in Berlin that Wolfgang Köhler (before he had to leave Germany) invented and conducted a famous experiment concerning the relation between the phonetic forms of two newly invented meaningless words and their potential psychological meanings. He drew two graphical representations at the blackboard, one with round smooth curves and one with sharp edges and acute angles, and he then asked his students which of these drawings was the meaning of the two words "Maluma" and "Takete", respectively. Not surprisingly, for his subjects "Maluma" sounded round and smooth, whereas "Takete" appeared to be sharp and acute. Thus Wolfgang Köhler got the expected answers.

I used phonetic reproductions of the two pure sound words "maluma" and "takete" to show that (i) new meaningless words can be invented and clearly communicated to an audience under normal noise conditions by just one single demonstrating utterance. Any student is able to give as many equivalent reproductions of "maluma" or "takete" as he is asked for. But I could also use Köhler's examples to introduce the concept of symbolically representing the category of a phonetic utterance. If, in a pairing situation, I pointed to simplified and stylized versions of Köhler's maluma- and takete-signs, i.e.:



I could also prove that (ii) - when in a pairing situation I had pointed to one of these signs and

---

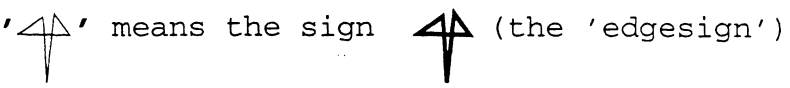
[4] For the first time presented in my inaugural speech on "Phonetik und Sprachliche Kommunikation" at the University of Munich and then further developed in Tillmann with Mansell 1980.

created at the same time just one single corresponding audible "maluma" or "takete" event - any student had understood this as an ostensive definition because he was able to reproduce the sound meaning of the graphically represented sound word.

In Köhler's experiment, the graphical representations were the heteronymic meanings of autonymically introduced sound words. In my modified application the meaning relations are exactly the other way round. I first introduced two signs as such (meaning themselves). To indicate the autonymic use of these signs I propose to use single quotes.

Thus

and



as a graphically given form or category. (We could even introduce names to identify these autonymically reproducible entities such as the 'roundsign' and the 'edgesign'.) After these signs were introduced as categorically reproducible categories I could - in the pairing situation described above - give them a new ***heteronymic*** meaning.

To indicate that a graphical representation stands for a phonetically reproducible category I propose to place it between double quotes. Thus the heteronymic meaning of " ◌ " is an autonymically reproducalble phonetic event, which (in Tillmann with Mansell 1980) had to be written down in an orthographic form, i.e. "maluma". But we have to keep in mind that the meaning of the expression "maluma" is just one concrete audible reproduction, autonymically demonstrating the category of itself. By only introducing the simple method of making this different use of single and double quotes, we can very easily express that 'ü' means the letter ü, whereas "ü" means a given speech sound produced by a speaker who is demonstrating the category of this phonetic event. Thus when reading in the context of this paper something like '"ü"', just say the soundword "ü".
Without quotation marks, maluma (or Maluma) can have many meanings depending on the context of situation where this word is used. In quotation marks it can have exactly only these two meanings: 'maluma' ≠ "maluma". The first expression means itself (a string of letters: 'm', 'a', 'l', 'u', 'm', 'a'), the second just one concrete demonstration of the phonetic form of " ◌ ".

We may either invent a new writing system for representing complex or elementary sound categories or just use one of the existing ones. In any case it is very helpful to have the distinction

between autonymic and heteronymic uses of sound signs. Thus we can define the more complex expressions

'  ' = 'maluma + takete'

'  ' = 'takete + maluma'

and, as soon as we introduce an ostensive definition and install the heteronymic meaning of '+' by producing "+" or "und", we even know, how "  " or "  " is to be reproduced phonetically.

We can also define the sound meanings of single letters such as 'm', 'a', 'l, 'u' by ostensively demonstrating "m", "a", "l" and "u"-events, respectively, and then represent the complexly defined -category by an analysing expression, such as

' ' = 'maluma'

If the phonetic meaning of 'm', 'a', 'l', 'u' has been effectively demonstrated by the respective autonymic reproductions we can even define quite complex new categories by writing what I call an analysing expression, such as

'  ' = 'mumulamulu'

When we are writing we normally are not interested in the letters nor in what they represent phonetically (and, accordingly, we don't have any need to employ single or double quotes). Thus in normal writing situations the user of an alphabetic writing system does not make any references to the written symbols (autonymically), nor to their (heteronymic) speech sound meanings, but to the normal heteronymic interpretations of the words of the language and what they allow the user to express semantically.

I. Considerations of principle

The history of speech science is a history of discoveries which show how little natural insights human speakers really have in what they are doing when successfully conducting speech acts. Many of the reasons for this lack of insight could be analysed in more detail, but there is in particular one which should be mentioned.

The processes of producing phonetic facts during natural speech acts run much too fast for any direct inspection. Even a trained phonetician looking at the midsagittal respresentation of a sound sequence such as "ich habe Stolke gesagt" (produced by a speaker of German) has to slow it down by a factor of about 5 in order to be able to observe the relevant details of the jaw and tongue movements. Obviously speech has to be such an automated and highly trained form of behavior and has to run at such high speed to serve the creation of semantic relations that are expressed by these fast actions.

2. Language words and sound words

When we compare the phonetic forms of certain words as they are uttered in normal speech acts (where they are used semantically by the speaker in their heteronymic meanings) and as we utter them in isolation in order to autonymically demonstrate, as clearly as necessary, their phonetically reproducible form, we observe all those differences that exist between pure sound words and real language words. Only in the second case do we see forms that are described in pronunciation dictionaries.

In Tillmann with Mansell (1980) we have proposed to specify the second type of phonetically very clear pronounced words as "alphabetically explicit". We also could show how important it is to understand that a phonetic transcription may have only one of two quite different truth value conditions. This depends on what is logically maintained by such a transcription. If the phonetic transcriber is interested in identifying the words of a given utterance he usually separates them by blanks and gives a narrow or broad transcription of their alphabetically explicit forms. In this first case of lexically representing a phonetically given utterance (we used the German term "wörtliche Darstellung" which cannot be literally translated into English) the transcriber is not really interested in specifying the actual phonetic form of this utterance. So the semantics of such a transcription is determined by the claim to maintain that the identifiable words are produced in isolation just as shown in the transcription.

On the other hand we get quite different truth value conditions if the semantics of the transcription is related to the phonetic forms that have actually been produced by a given speaker. As soon as we have to segment and annotate a naturally produced speech signal, the truth values of the resulting transcription are determined according to what can be really observed by a close inspection of the speech wave under auditory control.

It is exactly this kind of relations that exists between the phonetic forms of heteronymically used language words and the forms we observe under the condition of autonymic word demonstrations that we want to move to the center not only of phonetic speech research, but also of phonetic theories of speech acts. Therefore we would like to introduce a new terminology which reflects these two kinds of phonetic forms of words. We will continue to call the words of a language as they are used in their heteronymic meanings in normal speech by the speakers of that language ***language words***. But words which are identified by a graphical representation (which can be cited by the use of single quotes) and then are heteronymically interpreted by an autonymic phonetic event (which can be symbolically represented by the use of double quotes) will be strictly called ***sound words***. As phonetic objects, sound words can be identified semantically by simply looking at what in earlier publications of the present author has been called their observable articulatory content.

Sound words are always produced in isolation. They can be more or less complex or even quite elementary. Thus "♡" or "⌂" are more complex than "☺", "+", or "☤", and "90" is more complex than "9". We are, of course, interested in situations where the lexically given identifier of a language word can be represented by an analysing expression whose components are less complex sound words.

However, if

'99' = '9 + 90'

we would like to be able to specify, why (and in what phonetic details) "99" is different from "9", "+", "90" produced as a list of unconnected words so that

"99" = "9 + 90" ≠ "9" "+" " 90"

becomes true.

We are interested in answering the question how the sound words "9", "+", and "90" have to be changed in their phonetically given (and, to a certain degreee, also auditorily observable) articulatory content in order to become a given "99", produced, say, by the same speaker. And then we could look into a database of this speaker to find out how the sound word "99" is in agreement with the phonetic forms of this word when used as a language word. We could also ask the question whether we get at least a slightly different phonetic form if a given sound word such as "99" is just produced as a single language word in a neutral situation without any further context.

Another question is: where do we find sound words in real life? Pure sound words, both elementary and complex ones, are quite naturally produced as spontaneous demonstrations of equivalently reproducible phonetic events in two kinds of situation. When learning to speak a language the words of the language must be ostensively introduced by a teacher. But also when speakers and listeners of a language start to learn to read and write the words of their own language, sound words play a central role. Mothers produce sound words to demonstrate the category of a reproducible event to their children and teachers produce sound words to instruct their pupils in reading and writing.

Even phoneticians first have to learn to identify the audible qualities of the IPA-sound categories by being exposed to the respective elementary sound words; a good example here would be a proper demonstration of the tense and lax vowels of German in isolation. And here, my personal observation is that the students in our courses learn by themselves to give autonymically presented elementary sound words a heteronymic meaning by referring to their symbolic representations. Thus the vowel of the German word "Kind" is simply identified as the 'small capital i'. So we may have heteronymic and autonymic meaning relations between a sound word and its graphical representation in both directions.

Semantically, the interesting aspect of such cases is that the heteronymic meaning of these alphabetically elementary sound words is just the letter representing the sound word. Here, again, we get Köhlers meaning relation, where the words had a graphical meaning. So the heteronymic meaning of pure sound words would be nothing but their lexical notation such as, for every speaker of German who is not illiterate, the heteronymic meaning of the pure sound word "ü" seems to be primarily the letter 'ü'.

If this observation proves true we could even conclude that for the writers and speakers of a language a written word can take three different possible meanings, while a spoken word has

only two. The German word *Abc*, when written without any quotes, can simply be used as a synonym of the language word *alphabet*; if written as 'Abc' it just means this sequence of letters; and written as "Abc" it has the same meaning as "a-be-tse", pronounced as a sound word by a speaker of Standard German. But the speaker who is producing the word *Abc*, has only two options: the pure sound word means 'Abc', and, when he uses the language word *Abc*, he normaly will be referring to the Alphabet.

If there is not a third meaning of a language word (as in the case of k = kilo), the single letters in the lexicon such as *a, b, c*, etc., have obviously only two significant meanings, 'a' and "a", 'b' and "be", 'c' and "tse", etc.

3. Autonymic sound words and heteronymic language words in speech technology (SLP)

Our distinction between autonymic sound words and heteronymically used language words is also of interest for some of the actual problems of modern speech technology, in both major domains of SLP[5], i.e. automatic speech recognition and artificial speech synthesis. Even if pure sound words seem to be of little primary interest in any practical SLP-application, the relations we observe when comparing the phonetic forms of sound and language words may not be ignored, neither in speech recognition nor in speech synthesis. As the latter will be dealt with below in section 7, I restrict myself here to automatic recognition. In this case the relation between sound and language words are determined by the fact that sound words are very clearly articulated, while language words (at least in certain parts of an utterance) are much less clearly articulated. This will have to attract more scientific interest in the context of future SLP-research than it does today. The common aim of phoneticians and speech technologists must be to find methods for deciding whether a given piece of speech is clearly or less clearly articulated by a given speaker, and to what degree this would be the case.

My own career as a professional phonetician started in 1963 with a one year project on automatic word recognition. Together with the brillant technician of the phonetic institute at Bonn, Herr Rupprath, we created a hardware system consisting of hundreds of transistors, resistances, condensators, etc., with a microphone as input and, as output, 20 small lamps for indicating which one out of 10 italian cardinals ("zero", "uno", "due", ..., "nove") and 10 additional command words (such as "per", "diviso", "dacapo") had been spoken at each trial.

Fig.1 (in the web-version:) Picture of the first DAWID-System, 1964

The original DAWID-System, described at the ICA in Liège (cf. Tillmann et al. 1965), could only be so successful at that time because we made the prudent decision to take whole sound words as the central units to discriminate from each other, and not to try to reduce these com-

---

[5] It should be mentioned that the term SLP (acronym of 'Spoken Language Processing') has been proposed by Hiroje Fujisaki who initiated the ICSLP-Conferences as a common forum for speech science and speech technology. To my understanding phonetics and semantics are going to represent the two major parts of speech science in the SLP-domain.

plex sound words - "zero", "uno", etc. - to more elementary ones (such as "u", "e" or "i", "a", "z", etc). The acronym DAWID, by the way, stands for 'device for automatic word identification by discrimination' (which means that also here in the word DAWID the 'W(ord)' is in a central position).

Our second prudent decision was to concentrate on those acoustic properties of the speech waves whose measurements showed clear maxima for certain speech sounds, so that we could define a threshold for triggering discrete feature detectors. Such properties were, for instance, the frequency of the first formant, F1, the distance of the first and second formants, F2-F1, or the fricative zero-crossing density function. Thus we were even able to use elementary sound words such as "a", "e" or "i", "s" etc. for the testing of single feature detectors and for adjusting their thresholds. However, quite soon it became clear, that the proper triggering thresholds had to be set to a much lower, more sensitive level in order to get the expected feature detections in the case of complex sound words. So we discovered that the measure of certain "distintive features" of speech sounds depends to a great extent on how clearly the respective sounds are produced by the given speaker in a given sound or even language word.

The probability that there is not a great difference between the phonetic forms of corresponding sound and language words is rather high in the case of isolated word recogniton. (In the first DAWID-system we had great problems when we only told our speakers to produce pauses between isolated word productions. Indeed, it is not easy to instruct a speaker of Italian to produce pauses between the production of complex sound words which are longer than the silent intervals of the geminates within the words. In a quite naturally produced Italian "otto otto otto"-sequence the 'tt'-pauses can be about three times as long as the interword pauses.)

During natural speech production the phonetic facts created by the speaker have a clearness-measure that varies between the two extreme H&H-polarities of Lindblom's well-known Hyper-Hypo-dimension. It is not easy to see how this measure could be incorporated into todays HMM-technology. This measure is itself a variable of time, which may change its value from one syllable to the next, depending on factors such as the local tempo of a speech utterance. On the other hand it is quite clear that certain properties of clearly produced sound words may not vanish in any case. If a speaker of German wants to communicate the ownership of some money by uttering either "dies ist mein Geld" or "dies ist dein Geld" the listener must be able to observe the articulatory facts of the "mein"- or the "dein"-soundwords in order to understand which one of these two possessives has been used as a language word. Speech technology still has to conduct cosiderable research in order to solve the problem how to decide which sound word in the lexicon of a speaker was used by this speaker as a language word in a proper heteronymic function.

4. Random variation vs systematic modification

Why does phonetic knowledge not play a larger role in modern speech technology? One reason is certainly attributable to the fact, that purely statistical methods such as Hidden Markov Modelling or neural net computing produce much better results than so-called rule based expert systems. The situation is somehow self-contradictory, without any recognizable ways of effectively combining statistical and knowledge based methods. On the one hand rule based systems are simply much too powerful in two respects: they generate exploding sets of possi-

ble solutions that cannot be effectively computed in a reasonable time, and these sets also remain more or less empirically empty; and, theoretically derived forms are of little practical interest if there is not one real single utterance in a given database that falls under the specified category and could be taken as an instantiation of it. On the other hand, we must consider that even the largest databases of spoken utterances that have been collected up to now for the purpose of training and evaluating SLP-systems do not contain enough material to model most theoretically interesting cases. This is the dilemma of modern speech science.

The research aim in this situation can only be to find ways to treat phonetic variability (and to reduce purely statistical variation) by introducing the concept of systematic modification. A first step of reducing the amount of variability could simply consist in separating the data of individual speakers. Another way of reducing acoustic variablity could consist of relating the acoustic picture of sound and language words to the underlying articulatory processes which produce that picture, and then interpret the individual articulatory data by relating it to a generalized system. This could one day be obtained by means of a properly organized neural net which turns heard speech signals in some generalized newly articulated speech waves. A second step will be to look more closely at the prosodic form of speech productions analysing glottally controlled voice production and segmentally controlled sound articuation as a whole, i.e. in a totally intergrated way.[6] But the most important step would be to specify each lexically represented sound word with respect to its possible modifications that quite systematically have to take place as soon as this word, in a specified context of other words and of situation, is to be uttered by a given speaker as a real language word. I would like to illustrate this idea by some observations in Barbara Kühnert's dissertation, where she analyzed the t_k-assimilation of native speakers of German and English.

The German word "Blatt", produced as a sound word, shows a very clear final t-articulation. Used as a language word in a context like "Das Blatt kam von der Eiche" the t-behaviour of the tongue tip depends on whether this sequence is produced in normal or in fast speech. In very clear speech the language word still has a measurable EPG-t-contact; if this contact is lost, the electromagnetically measured movements of the front tongue may still show a whole scale of reductions. Barbara found in her data everything, (i) the tongue tip still contacting the alveolar region, (ii) the tongue tip moving almost all of the way up without getting into EPG-contact, (iii) the tongue moving half of the way, (iv) only a little bit, and (v) not at all. Thus there obviously is an H&H-continuum, that determines the behaviour of the tongue tip at the end of words with a final t-sound followed by a k-word.

That this articulatory reduction of the t-sound in a given situation of t_k-assimilation is not just random variability, but a very systematic modification depending on the situation, seems quite clear to me. In the final section of her dissertation, '7.5 Der Sprecher als Hörer' (p. 354 in FIPKM 34) Barbara Kühnert describes an experiment with two subjects (an English and a German phonetican) who had to judge their own reduced t_k-productions. The stimuli were

---

[6] I still have some hope that my early proposal to handle what I called "silbischer Ausprägungskode" (meaning the locally varying degree of 'well-articulatedness' of alphabetically specifyable soundsequences, the local degree of clearness of alphabetic soundstructure, the local tempo of syllable production, etc.) could consist in defining it as a computable function of intonation. Campbells prosodic concatenation system CHATR seems to offer a first verification of this basic idea.

short VCV-segments cut out from the respective reduced t_k- as well as from k_k-productions as control items.

The German subject (which was me) did quite well with his own reduced t_k-productions. My hypothesis is that in my productions there is - even if the t-movement of the tongue tip is reduced to zero - still a prosodic reflex of reduced segmental information that allowed me to reach a 100 % score in comparison with the control sequences. But this is a hypothesis that needs further investigation.

My hopeful conclusion concerning the possibility to separate the variability within the phonetic forms of language words into the components of statistical variation and systematic modification is strongly supported by this result. Only if in the two fast t_k- and k_k-versions there is still some systematic prosodic reflex of the missing t-movement the outcome of the experiment could not be explained as an artefact of something else.

II. Practical Applications

That sound words give us just the material we not only need, but can also simply take from any speaker of a language if we want to build up new utterances (of the same kind we observe in natural speech acts and which will be accepted by the speakers and listeners of that language as quite naturally produced utterances), is indeed a very challenging idea which I would like to further illustrate by describing the following three examples I borrow from ongoing projects in our institute.

5. The articulation of complex and elementary sound words

Probably the best way to understand which problems have to be solved in trying to approach our challenging goal is to look at some of our articulatory recordings. The data has been collected with our electromagnetic equipment (cf. Hoole 1996) which allows us to record up to ten fleshpoints from the speakers lips, jaw, and from about 5 cm of the front part of his or her tongue during normal and fast speech productions.

First of all, I should however mention that none of these articulatory projects is explicitly devoted to the ambitious goal which I'm talking about here. In none of our applications for receiving our project-grants in articulatory phonetics has the role of the sound word as a central phonetic unit actually been mentioned. Thus it should be clear that in our research we are still dealing with much more specific questions that can be answered by analysing the data itself (without trying to modify them in a proper way).[7] The following picture illustrates the movement of the front part of the tongue during a normal and a fast production of the utterance

---

[7] On this occasion I should specially thank the German Research Council DFG for sustaining our work by grants Ti 69/29 (articulation of the German vowel system), /30 (development of 3-d-EMA, with only one 'M'), /31 (our contribution to the 'DFG-Schwerpunkt Sprachproduktion'). The work in these projects has been or is done by Phil Hoole, Barbara Kühnert, Andreas Zierdt, Christian Kroos, Christine Mooshammer, and Anja Geumann.

"Der nette Schotte hat eine schwarze Socke verloren" (produced by three speakers).

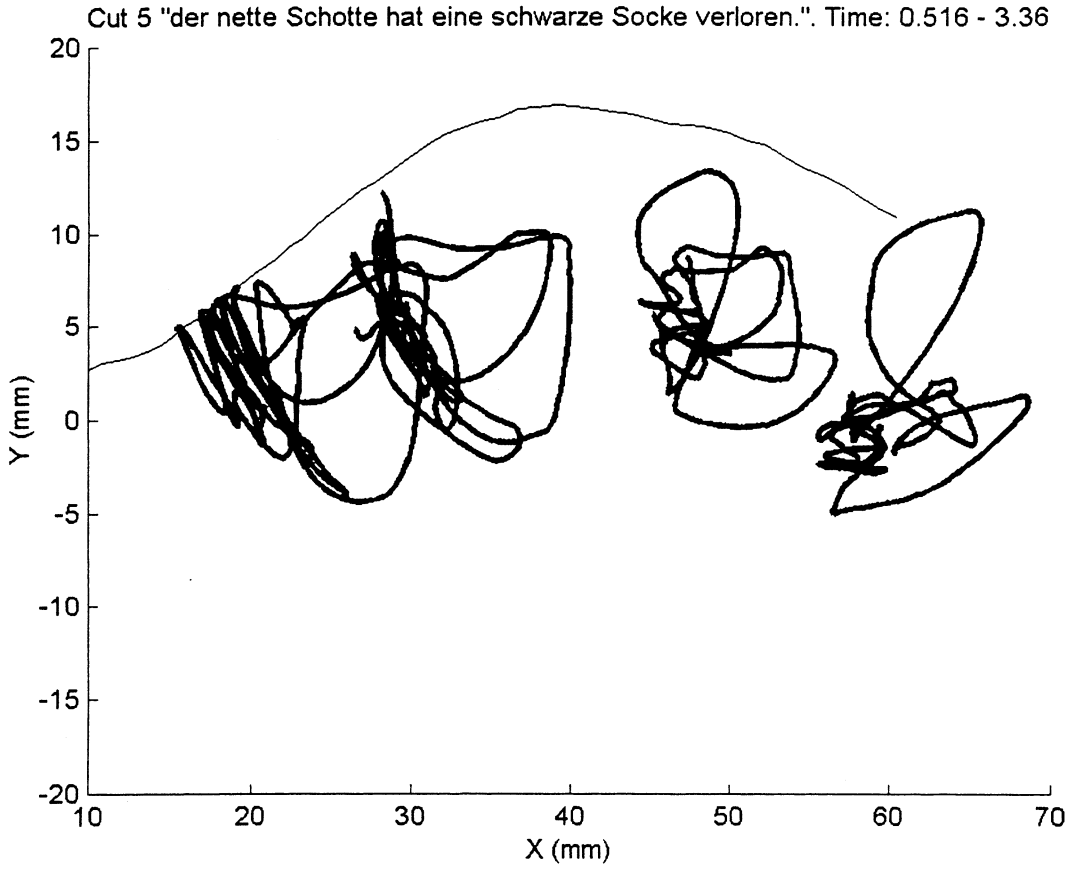


Fig. 2 (cf. the color-versions of this example on the web)

Without listening to the acoustic results of these speech movements no trained phonetician will be able to decide which complex or elementary sound words have been transformed here into the language words by the three speakers who were reading the prompting word sequence in two different speeds.

To get a better insight into the obviously very complex processes of transforming isolated sound words into a sequence of real language words I would like to propose to begin with elementary sound words such as "m", "a", "l" etc. and then to see, how these have to be modified in order to form a lexically given "maluma".

As everybody knows, things are a little bit more complicated than this would imply. We only have to mention the most elementary processes of coarticulation and assimilation at the lowest level to be reminded of this. A good example is the mid vowel tongue position of the tense and lax German vowels in dependency of the CVC-context as in the case of C = p/t/k.

On the other hand we have to consider that certain alphabetically represented sound words are not really elementary. Sounds such as [k], [g], [t], [d], or [p] and [b] cannot be demonstrated as elementary sound words without also producing a voiced or voiceless vowel.

A much more complicated "elementary" situation is given in the case of the so-called German

vowel opposition, which is observed in pairs such as "Miete, Mitte", "Mühle, Müller". No phonetically untrained speaker of German is able to demonstrate the vowel of "Mitte" or "Müller" as an isolated single-vowel sound word. The Standard German sound system contains only those elementary sound words which are represented in the lexicon as "i", "e", "ä", "ü", "ö", "a", "o", and "u" which are listed in the Duden for instance by their respective letters.

Here, by the way, I would like to confess that (unlike most of my colleagues today) I totally agree, as a phonetician, with Theo Vennemann's analysis, which clearly shows that from a phonological point of view the German vowel opposition can not be reduced to an elementary sound word opposition. The distinction is certainly a prosodic one and can only be dealt with in closest connexion to the syllable structure of German[8]. I would even propose to introduce the concept of prosodic modification in its strongest sense. So we simply take, for instance, a generalized articulation of demonstrating an elementary German "o"-word

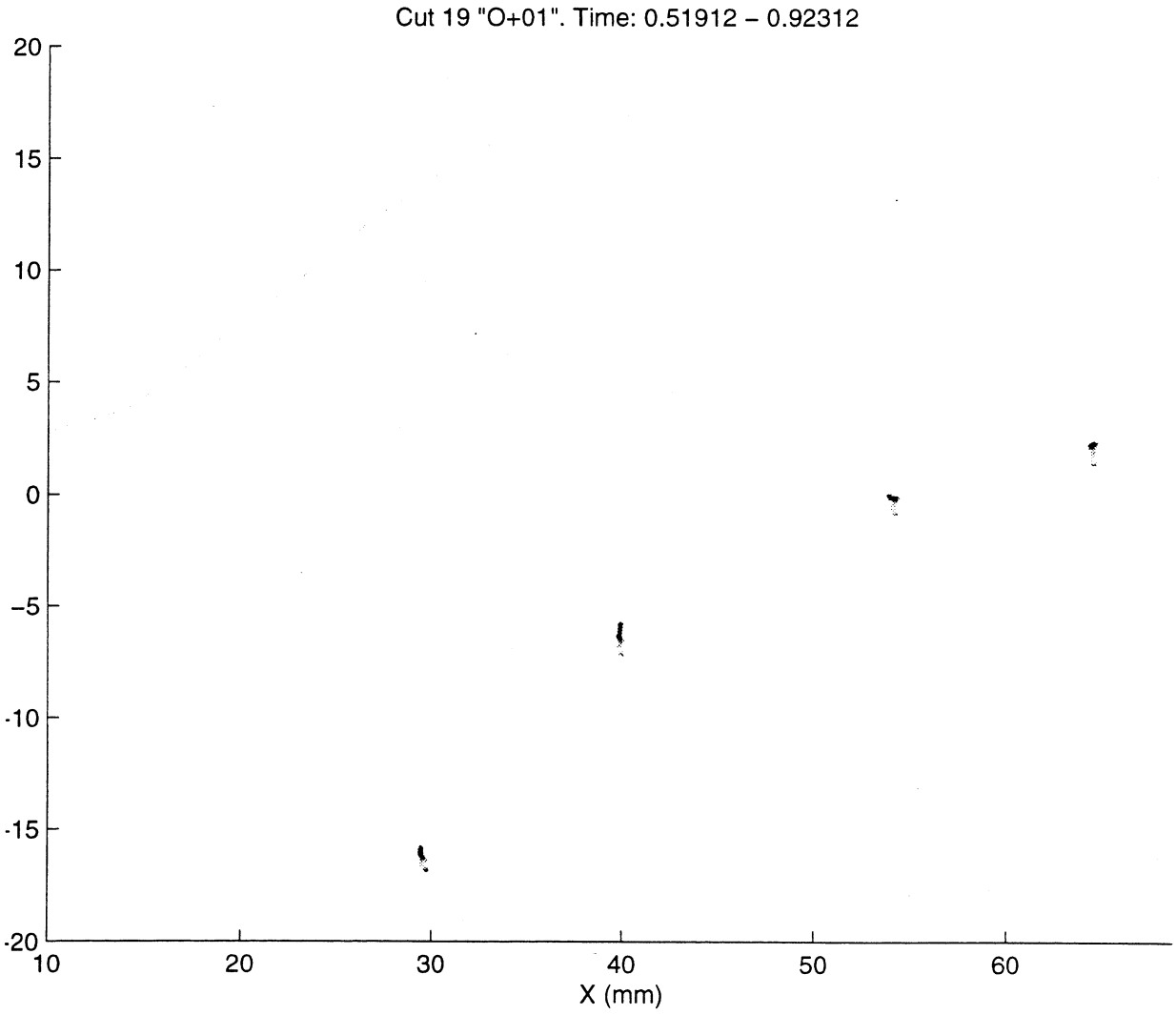


Fig. 3 (cf. the color-versions of this example on the web)

[8] See also T. Becker 1998 and D. Restle 1998

not only for modifying it into a complex sound word such as "gepope", "getote", "gekoke" (in two different tempos):

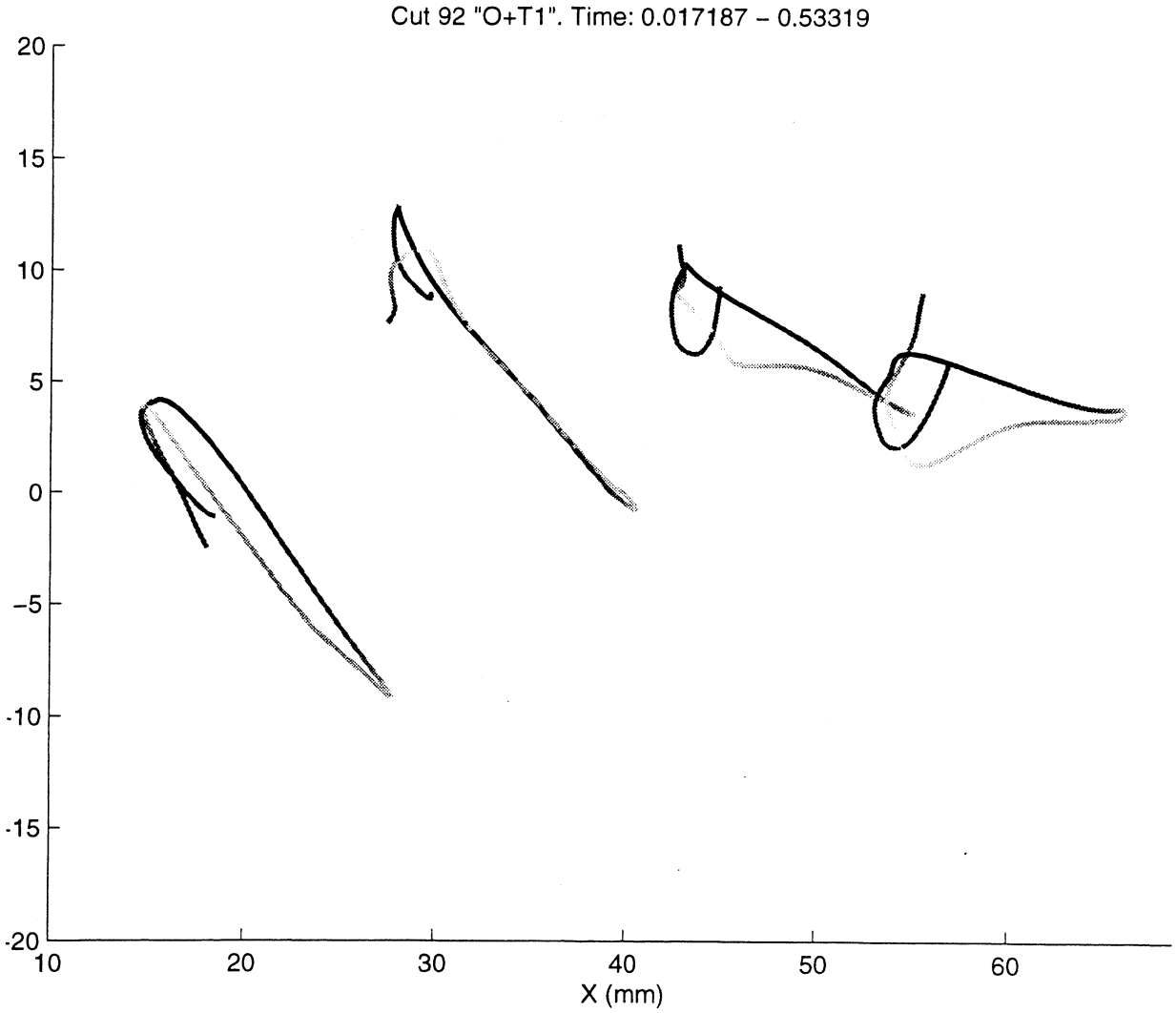


Fig. 4 (cf. the color-versions of this example on the web)

but also to systematically change it into the "gepoppe", "getotte", and "gekocke" sound words (also in the respective tempo versions).

(see Fig. 5 on next page)

Everybody can see which kind of transformations have to be taken into acount in order to achieve our ambitious aim.

When we have arrived at a theory which specifies all necessary algorithms for the linear and nonlinear interpolation between elementary sound words and their complex counterparts in the sound word lexicon of a speaker, we shall also be in the position to do the next step of developing the theory that further transforms complex sound words into proper language words (so that they look just as those in *Fig. 2*, above).

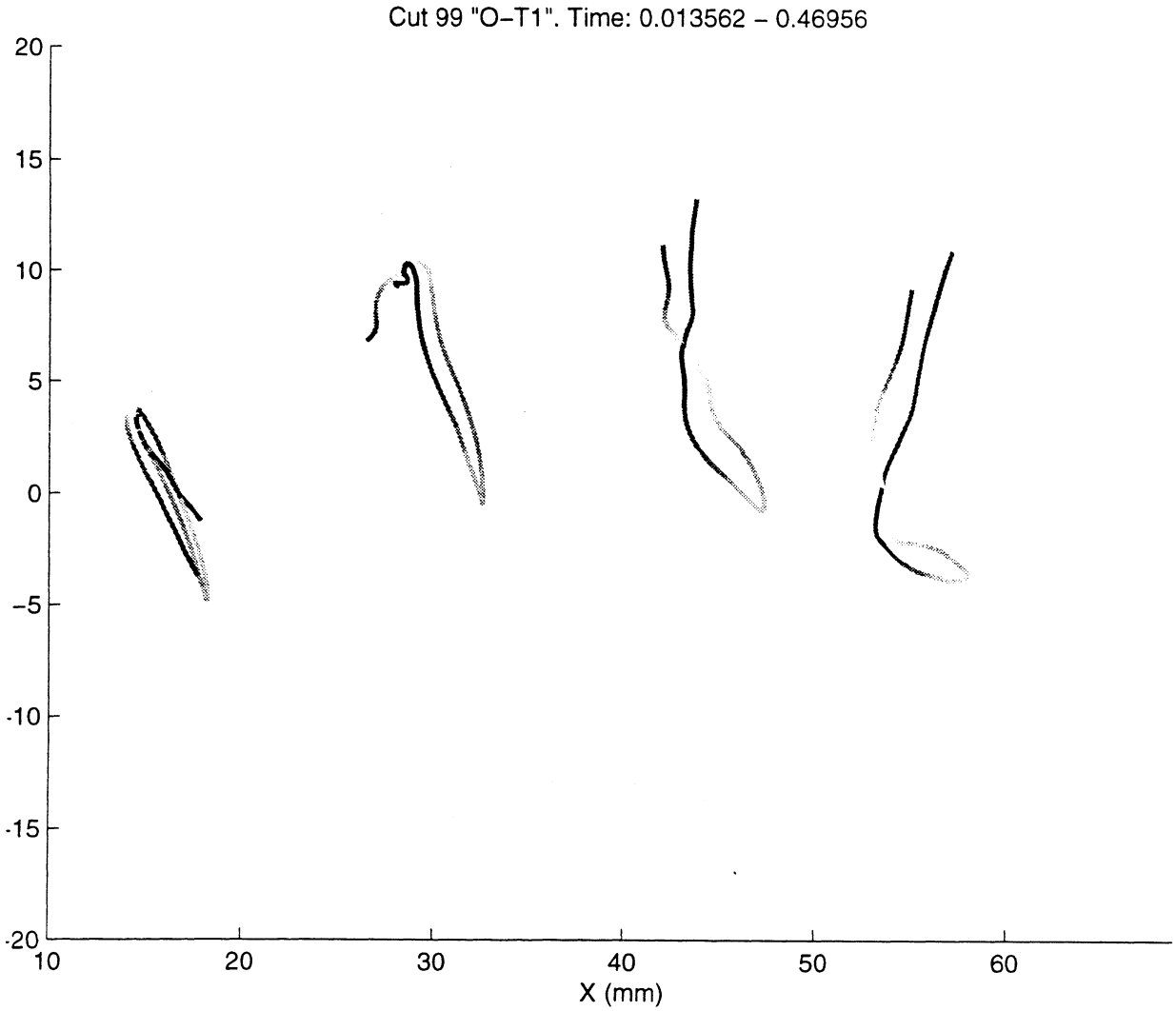


Fig. 5 (cf. the color-versions of this example on the web)

The details of such an articulatory theory of producing natural utterances for conducting real speech acts will be extremely complicated (even if most of the work could probably be done by neural net computing methods), but the basic idea remains quite simple, that is of taking the sound words of a language as the central units. We are convinced that in the case of speech synthesis it is much easier to take something already given and change it into something more or less different than to create something from nothing.

The words to be changed must not only be stored as data, but also categorically specified within the lexicon of either a given individual or a computed generalized speaker of the language in such a way that all possible modifications of the given sound words as well as all necessary transformations for producing proper language words are sufficiently determined. Here particularly phonologists can do a very helpful job if we phoneticians supply them with an

experimental artificial speaker.[9] For instance, the sound words of Standard German must be specified with respect to which syllables are "tone-syllables" in the sense of Thomas Becker and so must be prosodically processed with respect to the vowel opposition of German, and which are not and are to be treated differently.

6. The central role of the word in the preparation of databases for SLP-technologies

The second excample should only be mentioned here, because Christoph Draxler is giving a whole talk just on this topic. Let me shortly address some of the essential viewpoints.

First of all I want to point out that in PHONDAT (and later in VERBMOBIL) we had a close collaboration between the phonetic institutes of Kiel, Bonn, and Munich, and especially Klaus Kohler and the present author took great care that the CRIL-conventions of the IPA were strictly applied in the organisation of all our speech data collection. In this context we introduced the term 'canonic word form' which to my personal understanding could also be used to refer to a clear lexical representation of the corresponding sound word as described in a narrow SAMPA-notation. Christoph Draxler will come back to this term because he has translated it into a PROLOG-predicate which plays the central role as a unit for organizing the databases of PHONDAT (and VERBMOBIL).

CRIL is the acronym of 'computer representation of individual languages'. The conventions have been defined at the Kiel convention of the IPA and say that there should be at least three levels of symbolic annotations for any given speech signal: (i) an orthographic representation (if possible) to guarentee the identification of the lexical entities produced as language words within the given utterance; (ii) a broad phonemically oriented notation of a possible citation form, which corresponds to our complex sound words; (iii) a narrow transcription of what has actually been pronounced by the given speaker which corresponds to what I refer to as the phonetic form of the actually produced language word. This third level is directly related to the speech signal and indirectly related to the canonic forms by systematically indicating insertions, substitutions, and deletions of sound segments. What is still missing (and has not yet found a regulation with respect to some CRIL-convention, but will be certainly needed in the near future) is a method of specifying the degree of articulatory clearness ('Wohlartikuliertheit'). My prediction is that these prosodic components of a given utterance will be handled in the near future by some automatically derivable measures which are directly related to the phonetic properties of the speech signal at non-symbolic level.

Using speech data (which had been labeled in Kiel and Munich according to the CRIL-conventions) we have developed the MAUS-system that automatically segments and labels spontaneously produced speech utterances under the condition that (i) the sequence of language words within this utterance is orthographically specified, and that (ii) a canonic representation of the corresponding sound words can be looked up in some kind of a pronunciation dictionary containing a regularly defined canonic word form.

In this context, any word of a given language (or the ideolects of the speakers of this language) could be seen as a theoretically specifiable object that contains all possible systematic modifications in relation to its canonic sound form (as can be seen in the graphs of Florian Schiel's

[9] See also Kohler 1991.

contribution to this workshop). The MAUS-system works only under the condition that the canonic form is known and we can specify a corresponding theoretical object containing all possible segmental modifications of the sound word when used as a language word. In the analyzed speech wave, these can be empirically verified with respect to the actually given form of the language word as it was uttered by the speaker. In extreme cases, as we all know, the phonetic form of a sound word when used as a language word can be reduced to zero.

Florian Schiel will describe the MAUS-system in his contribution to this workshop in more detail. I may restrict myself to say only two more things. I wish to confess that I am proud, indeed, about the fact that our idea of verifying the phonetic forms of language words given only their descriptions as sound words proved to work so beautifully, even in its present state (which will be further developed of course, in the near future). Secondly, I wish to emphasize again that it is only because we took the word as a central phonetic unit we were in a position where we could start to automatically collect the knowledge we need for developing a CPT of German[10].

3. PHD: From sound words to connected speech

My last example is almost future music, because many components of this new project are still in planning stages. The acronym PHD stands for 'parametric high definition speech synthesis'. Together with Hartmut Pfitzinger and Kurt Kotten I have begun to develop an experimental system that will allow us to systematically modify sound words produced by a given speaker and transform these then into language words acoustically.

Several years ago, in a DFG-Schwerpunkt (Sprachpsychologie) we designed and realized a package of DSP-programs to define a continuum between different speakers uttering the same sequences of language words. The system interpolated between presegmented parts of these utterances and produced a set of stimuli for conducting experiments on categorical perception of speaker identities (cf. Tillmann et al. 1984). In the new PHD-project we are, at least at the beginning, less interested in interpolating (and extrapolating) inter-individually between 'the same utterances' of different speakers, but between different sound forms of the same words produced by the same speaker. We are developing these methods of intra-individual interpolation between given utterances and therby producing acoustically new utterances because we believe that an experimental work bench of this type is needed to learn more about what it means that an utterance of a word can possess a variable degree of clearity. How do we have to reduce phonetic properties and time durations of a given sound word to transform it into a realistic language word of the particular speaker in the proper contexts? Thus, quite differently from our ideas in the articulatory oriented example above, in the PHD-project we are less interested in combining elementary sound words to the complex sound words, but in reducing complex sound words into realistic language words.

The PHD-system is not designed as a typical text-to-speech system, but (in a certain sense) as a 'language_word_system', and we are designing this system mainly for conducting experimental investigations concerning the central question that I think is the most important one

[10] The goal of developing a strictly empirically based 'complete phonetic theory' of a spoken language has been proposed by Pompino-Maschall and the present author in our contribution to the EUROSPEECH conference in Berlin, 1993.

that phonetic speech research has to answer in the near future: How can we theoretically and also practically relate the phonetic forms of language words to the clearly defined properties of the corresponding sound words as they are autonymically demonstrated by the speaker.

What we do have to change when going from one speaker to another speaker of the same language is quite another question that can only be answered as soon as we know what these speakers already do by themselves when they are changing their data intra-individually from one language word to another one and in relation to the given sound word.

8. Concluding remarks on simplicity and complexity

One of the traditional goals of linguistics has always been to reduce the complexity of phonetically given utterances to a simple underlying grammatical representation. The only way, I think, I could agree which this traditional goal would be to reduce any regularly produced speech utterances to the words which are contained in those utterances, and representing these words for any given utterance by a lexically specified object which is the set of all its possible modifications under defined conditions. Such a theory will be a rather complex one, but it will be governed by the clear and simple idea that the word is the central unit of speech production and speech perception.

There is still quite another aspect concerning the complexity or simplicity of speech processes. All phonetic facts which are to be theoretically modelled by a phonetic object incorporating a complete picture of the potential segmental and prosodic sound structures of each single word of a spoken language, are facts given at the periphery of the speaking nervous systems. The speakers and listeners are able to - and have to be able to - directly observe them as they appear during each act of speech. If we call these facts in comparison with their linguistic descriptions complex, we must, on the other hand, understand that they are (relatively) extremly simple - if we only compare them with those (really extremely) complex processes that have to take place within our human nervous systems during any act of speech, be it an act of demonstrating an elementary sound word, say "o", or conducting a real act of speech.

If we only look at the articulatory content of sound and language words I believe that, in the near future, we will begin to understand that the particular phonetic form which a word takes in a given utterance, is nothing else but a computable prosodic function of what the speaker wants to express semantically.

References

Becker, T.: Das Vokalsystem der deutschen Standardsprache. Frankfurt 1998

Draxler, Ch.: Database systems for spoken language corpora. (this volume)

Hoole, P.: Theoretische und methodische Grundlagen der Artikulationsanalyse in der experimentellen Phonetik. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM) 34, 3-174, 1996

Kohler, K.: Prosody in speech synthesis: the interplay between basic research and TTS application. Journal of Phonetics 19, 121-38, 1991

Kühnert, B.: Die alveolare-velare Assimilation bei Sprechern des Deutschen und Englischen: Kinematische und perzeptive Grundlagen. FIPKM 34, 175-392, 1996

Paul, H.: Prinzipien der Sprachgeschichte. 1898. Zit. nach: 5. Aufl., Tübingen 1920

Restle, D.: Silbenschnitt - Quantität - Kopplung. Phil.Diss., München 1998

Schiel, F., and Kipp, A.: Probabilistic analysis of pronunciation with 'MAUS'. (this volume)

Tillmann, H. G.: Das phonetische Silbenproblem. Phil. Diss., Bonn 1964

Tillmann, H. G.: Eight main differences between collections of written and spoken language data. FIPKM 35,139-144, 1997

Tillmann, H. G., Heike, G., Schnelle, H., Ungeheuer, G.: DAWID - Ein Beitrag zur automatischen Spracherkennung. Paper A 12, 5th ICA, Liège 1965

Tillmann, H. G., mit Mansell, P.: Phonetik. Stuttgart 1980

Tillmann, H. G., Schiefer, L., and Pompino-Marschall, B.: Categorical perception of speaker identity. Proc. 10th ICPhS, 443-449, 1984

Tillmann, H. G., and Pompino-Marschall, B.: Theoretical principles concerning segmentation, labelling, and levels of categorical annotation for spoken language database systems. EUROSPEECH'93, 1691-1694, Berlin 1993

The disappearance of words in connected speech

Klaus J. Kohler
Institut für Phonetik und digitale Sprachverarbeitung
der Christian-Albrechts-Universität zu Kiel

1 Introduction

The category of the word is well established in meta-language pursuits, especially in linguistics. It is the basis for the development, over centuries, of the methods of lexicography, which have produced various types of lexica, including pronunciation dictionaries. The latter take the concept of the canonical phonetic representations of word citation forms in a language as their point of departure. At least some of them also list phonetic variants, which may occur in different utterance contexts (phonetic environments, levels of style). The pronunciation dictionaries presently available for various languages, differ enormously in the extent of taking the phonetic variability of words into account. At one end of the scale we find the DUDEN lexicon for German [1], which does not provide any contextual variants at all; at the other end are the English reference works by Jones/ Gimson/Roach [9] and Wells [37], which give detailed information on phonetic variation of words in utterances. WDA (*Wörterbuch der deutschen Aussprache*) [39] is located in between, but closer to DUDEN.

This concept of word pronunciations and especially of canonical citation forms is built on the idea of the independent existence of individual words in an utterance, which in turn results in the assumption that words are units that can be defined and delimited phonetically. The word thus also constituted the frame of reference for the development of phonology: phonemes are the sound units that differentiate words, and boundary signals mark their beginnings and ends phonetically. As a consequence of this focus on word phonology the level of segmental sentence or utterance phonology was largely excluded from the study of sound patterns in languages.

Typical, often cited examples of phonetic markers for morphological structure are the palatal fricative in German *"Frauchen"* (noun *"Frau"* + diminutive *"-chen"*) vs. the velar fricative in *"rauchen"* (stem *"rauch-"* + verbal ending *"-en"*) and the dark lateral in (Southern British) English *"coolish"* (adjective *"cool"* + loose derivative *"-ish"*) vs. the clear one in *"foolish"* (stem *"fool-"* + integrated suffix *"-ish"*). Lehiste's classic study of 'internal open juncture' of 1960 [25] also belongs to this field of word phonology (e.g. *"nitrate"* vs. *"night-rate"* or *"a name"* vs. *"an aim"*).

Nevertheless it has been known from historical linguistics for a long time that potential phonetic boundary markers for word separation may be ignored at any time. In this connection we may compare *"Natter"* and *"Otter"* in German, and refer to English *"adder"*, *"apron"* (as against *"napkin"*), or - with the opposite direction of sound change - *"newt"* (besides *"eft"*), *"nickname"* (as against *"eke"*): in all these cases the sequence of indefinite article and noun receives a new phonetic parsing. If modern linguistics had been dominated less by English and more by e.g. French the search for phonetic word indices would probably never have arisen. French, as a syllable-timed language with no lexical stress, lacks the phonetic marking of word units to a far greater extent than the stress-timed languages English and German, and consequently word puns abound. Here is a typical example:
"De quelle couleur est toujours un coffre-fort quand on le vide?"
"Il est tout vert". - "Il est ouvert."

This word orientation also determined a phonetic research paradigm which highlighted the word frame in experimental analysis still further, e.g. by the use of systematically varied nonsense words or of word contrasts in a constant utterance environment of the type *"It's a .... (Say...) again."* Although it is a reasonable assumption that the word is a language reality, at least for speakers, it reaches different degrees of awareness according to the demands of the communication situation, i.e. the word as a unit of speech will be particularly prominent in data collection under lab conditions, but far less so in spontaneous interchange. An investigation of the former kind uncovers coarticulatory effects and assimilations that stress the integrity of word units much more strongly, e.g. in the lack of complete labial/dorsal assimilation of coronal plosives and nasals at word boundaries, as in English *"hat pin"* or in German *"Schrottplatz"*.

The experiments by Nolan [27] and Kühnert [24] using EGP and EMA techniques, respectively, are cases in point. Having investigated apparent place assimilations Nolan went one step further in the interpretation of his data by proposing that differences in lexical phonological form always result in distinct articulatory gestures, even if overlapped and/or reduced or not discernible in the instrumental record. This is the complete reification of the phonetic word. But there is a good deal of evidence that the word boundary can be overridden in such cases, resulting in complete assimilation, especially frequent in, but not limited to, the reduction of function words, as in German *"mit dem"* with **[mɪpm], [mɪm̪m̪]** or **[mɪm]**, besides **[mɪtm]**.

The few examples quoted so far will have demonstrated that words may be identifiable as phonetic units but that they may also lose this phonetic identity, either by the change or the disappearance of boundary signals or by the entire fusion with other words. The phonetic manifestation of words thus oscillates on a scale from distinct separation to complete integration. The conditions for this phonetic variability of word identity depend on a number of factors:

- the general articulatory strategies in human language
- the individual language concerned
- the word class as well as the morphological and syntactic structures
- sentence accent, position in the utterance and general phonetic environment
- and, above all, the demands of the communicative situation as regards the balance between articulatory ease and auditory distinctivity, which is adjusted differently for different speaking styles - lab speech, read speech, spontaneous dialogue etc.

In order to be able to come to grips with this question of the word as a phonetic unit it is essential to go beyond the prevalent pattern of word phonology and move on to a consideration of the sound structures above the word at the sentence and utterance levels. This has been a focus of research at IPDS Kiel since the early 1970's and was mirrored in a German Research Council funded International Symposium on "Sound Patterns of Connected Speech", organised at Kiel in June 1996 [36]. As regards German, there is now a suffiently large, phonetically annotated acoustic data base of read and spontaneous speech, of altogether 70,000 running words, completely transcribed segmentally and in part also with prosodic labels: 'The Kiel Corpus of Read/Spontaneous Speech' on four CD-ROMs so far [4,5,6,7,21]. Together with a data bank environment and appropriate search as well as analysis tools it provides the necessary facilities [8,20,22,28] for large-scale corpus studies of connected speech processes in German [3,10,12,13,14,15,16,23,29,30,33,34]. A research grant from the German Research Council that has recently been allocated to IPDS for this type of investigation will allow us to exploit these speech resources more fully.

Moreover, this Conference at ZAS testifies to an ever growing awareness of the need for phonetic analysis of connected speech, and the organisers are to be congratulated for their initiative to run it by the side of a linguistics meeting on the **phonology** of the word. Therefore I am particularly grateful to them for giving me the opportunity to hold up the flag for phonetics by inviting me to speak to you here today. Since the obliteration of words in speech is more interesting than their preservation I have chosen the former as my subject.

2 The disappearance of words as delimitable units in speech production

2.1 Function words: from separation to integration

2.1.1 Disappearance of syntagmatic and paradigmatic phonetic word distinctions

The interference with phonetic word identity is particularly frequent in (sequences of) function words, e.g. in German

"Hast du einen Moment Zeit?" **[haspm mom'ɛn ts'aɪt]**

"Hast du den Bericht über die letzte Sitzung endlich geschrieben?" **[haspm bəʁ'ɪçt]**

The same phonetic form **[m]** in the strongly reduced sequence of three function words *"hast du einen/den"* can be uniquely identified with *"einen"* in one context and with *"den"* in another, although the solely remaining nasal (with labial adjustment to the following consonant) can no longer trigger the phonetic identification of the word. The separation of these words is further hampered when instead of **[pm]** a glottalized nasal **[m̰]** is produced, which signals the article and the plosive residual of *"du"* at the same time.

But the reduction can go further and eliminate all traces of *"du"* in **[hasm̩ mom'ɛn ts'aɪt]**, with a syllabic nasal, which may in turn follow the general German geminate reduction, especially in unstressed position and fast speaking rate, resulting in **[has mom'ɛn ts'aɪt]**, where the reflex of *"einen"* has also disappeared in the phonetic manifestation. The verbal paradigm as well as the idiomatic phrasing make the decoding of the intended meaning of the utterance unique, and the listener therefore does not depend on the signal detection of every word.

2.1.2 Emergence of new words through syntagmatic fusion

The disappearance of words in context is not restricted to the loss of all phonetic traces but may also take the form of the appearance of new lexical items through the complete fusion of others. This is particularly common for prepositions + articles, as in French *"au"*, *"du"* or German *"im"*, *"ins"*, *"zum"*, *"zur"*. In today's usage, German *"er geht zur Schule"* and *"er geht zu der Schule"*, *"er kommt zum Schluß"* und *"er kommt zu dem Schluß"* have different meanings although both forms are historically related on a scale of articulatory reduction.

Similarly, subject pronouns in enclitic position to function verbs form a scale from separation into two items to fusion into a single new one in e.g. German *"haben wir"*, *"sind wir"*, *"hat er"*, *"habt ihr"*:

[haːbən viːɐ] [zɪnt viːɐ] [hat ʔeːɐ] [hapt ʔiːɐ]
[ha(ː)ɱ vɪɐ] [zɪɱ(pᵛ) vɪɐ] [hat (ʔ)ɛɐ] [hapt (ʔ)ɪɐ]
[ha(ː)ɱ vɐ] [zɪɱ vɐ] [hatʰ ɐ] [haptʰ ɐ]
[haɱ ɐ] [zɪɱ ɐ] [hat ɐ] [hapt ɐ]
[hamɐ] [zɪmɐ] [hadɐ] [haḇḍɐ].

The same subject pronouns in proclitic position and the indirect object *"ihr"* (in e.g. *"er hat ihr geholfen"*) reduce less, the possessive pronoun *"ihr"* (in e.g. *"sie hat ihr Kleid gewaschen"*) least:

in these cases fusion does not occur. So the disappearance of words in context and the appearance of new ones is not only situationally determined but also morphologically and syntactically.

2.1.3 Incomplete word fusion

A third type of the integration of words is illustrated by some of the reduced phonetic variants in *"die können wir uns abholen"* and *"die könnten wir uns abholen"* vs. *"die können uns abholen"* **[dɪ kœɱʸ ɱʸ ʊnʸs]** and **[dɪ kœɱʸ ɱʸ ʊnʸs]** vs. **[dɪ kœnn ʊns].**
Here words neither disappear without trace nor are they fused to new units: on the one hand, the sequential articulatory movements are greatly reduced, but, on the other hand, phonetic components of velarization, glottalization, nasalization etc. are kept as long residual traces of the eliminated elements overlaying the remaining ones. In these instances the tendency towards integration by articulatory fusion is counteracted by the opposite tendency to maintain the phonetic identity of the word through **articulatory prosodies** [11].

An extreme case of this is found in the series of four function words of
"nun wollen wir mal kucken", for which the Kiel Corpus of Spontaneous Speech provides
[nũː õnᵘ ɛ̃ ɱa kʰˈʊkŋ] (OLVg122a009); see spectrogram in [36], p. 2.

The four initial function words of the sentence *"**da hat er auch** keine Zeit"* (with the sentence accent on *"Zeit"*), which are clearly separated in the precise pronunciation **[daː hat ʔeːɐ ʔaʊχ]**, may reach the stage of complete fusion in **[daːdɐ ɑχ]**, where **[daːdɐ]** approaches a new lexical (phrasal) item as part of a paradigm **[daːβɪç] [daːstə] [daːtsə] [daːmɐ] [daːptɐ] [daːmzə]**. But the componential element of breathy voice control may be kept as a residue of /h/, superimposed on the vowel of **[daː]** preserving the identity of the word *"haben"*, as in the following example from the Kiel Corpus of Spontaneous Speech: *"**da haben Sie auch wieder** recht natürlich"* **(HAHg071a019)**; see spectrogram in [34], p. 140.

2.2 Componential residues of segmental deletions

In all these residue cases the componential features have to be represented in a phonetic transcription, even if it is basically segmental, because they mark phonological contrasts at the level above the word. In our labelling system in the Kiel Corpus, we have adopted the symbol **-MA**, inserted into the canonical transcription before symbolically deleted segments [21]. Its use may be illustrated by the following example (see spectrogram in [34], p. 157 and [11]).

TIS071a004		***"wahrscheinlich ein bißchen"***
canonical	**SAMPA**	**v a:6 #S 'aI n l I C Q aI n+ b ' I s C @ n**
variant	**SAMPA**	**v a:6 #S 'aI n -MA l- I- C- Q- aI- n-m+ b -h ' I s C @- n**
	IPA	**[vɑʃaɪɲ m bɪs͡ççn]**

- The syllable **l I C** is characterized by palatality, i.e. by a high elevation of the tongue dorsum, which is obvious for **I C** but also applies to the clear (palatalized) **l**. **I** before **C** is, moreover, produced with a higher tongue position than before non-palatal consonants, e.g. in the suffix *"-nis"*. So the difference between **I** and **C** is one of vibrating and open glottis with very similar tongue heights; these phonation differences together with similar oral strictures generate laminal versus turbulant airflow at the tongue-palate opening, resulting in approximant and fricative articulation, respectively.
- **l** and **I** are articulatory opposites in their central and side tongue-palate contacts, which puts high demands on the execution of the speech gesture chaining.

- The tongue tip/blade gesture is subordinated to tongue dorsum and lip movements; therefore, under these sequential constraints, the palatalized **l** loses its central coronal contact in the dental/alveolar area by adjusting to the purely dorsal gesture of **I**. This is found generally in the suffix *"-lich"*, e.g. in *"selbstverständlich"*, *"natürlich"*, particularly when the words are unstressed and non-final in the utterance.
- In unstressed syllables all articulatory gestures are probably reduced in their magnitude, including subglottal pressure and glottal opening for **C**. The result is the transformation to an approximant with the possibility of voicing: **j**.
- The dorsality of the reduced final syllable may then also be extended to the preceding nasal, due to the higher rating of dorsum over tip/blade gestures in articulatory sequencing, resulting in a palatal.
- With the desynchronization of velic movement, especially between two nasals, i.e. before **m**, which originates from the reduction and assimilation of *"ein bißchen"*, the dorsal approximant is nasalized as well.
- If the closing of the lips for **m** occurs early enough there will not be an approximant stricture between the nasal of *"wahrscheinlich"* and the nasal **m**.
- So we end up with the pronunciation found in the spontaneous speech example as a consequence of natural constraints on articulatory gestures: a componential residue of palatality remains although segmental units corresponding to a canonical form can no longer be separated.

The application of **MA** is particularly important in the case of the deletion of a vowel as a voiced sonorant stretch at the segmental level. This will now be discussed with reference to variants of the word *"vielleicht"* in the Kiel Corpus of Spontaneous Speech. Among the high-vowel elisions, this word supplies a very high incidence of **MA** markings [3,16]. So the analysis of the phonetic realisations of this item will be particularly informative from the point of view of spontaneous speech motor control. The following labellings (in SAMPA notation [38]) of the first, unstressed syllable will be considered: **f I l**, **f -MA I- l**, **f I- l** (see spectrograms in [3], pp. 122-127).

The gestural components that make up this speech unit are a labiodental stricture, a high front dorsal tongue position, a coronal closure with simultaneous lateral opening and a glottal abduction-adduction sequence. The precise temporal coordination of all these constituents is highly variable and continuous rather than discrete. In the hyper-version of the utterance the high dorsal position is maintained after the labiodental release, in turn followed by the tongue tip/blade and side gestures, and synchronised with voice onset. Deviations from this organization in the corpus data are:

- voice onset is delayed in relation to the labiodental release along a scale up to complete devoicing of the vowel;
- the onset of the coronal gesture is advanced in relation to the labiodental release along a scale up to complete disappearance of a separate vowel element; the syllable timing may otherwise remain unchanged (resulting in a syllabic lateral) or get shortened concomitantly (producing a non-syllabic lateral);
- the advanced timing of the coronal gesture may be combined with voice onset delay, again along a scale, resulting in more or less devoiced laterals;
- the high, front dorsal tongue movement may be kept in spite of the early coronal timing (resulting in palatalization within the **f l** cluster, particularly after an immediately preceding front tongue elevation, e.g. in the word *"nich(t)"*), or there may be early coarticulation with the diphthong **aI** onset of the subsequent syllable (and thus coalescence with the word-initial cluster **f l**).

All the realizations of *"vielleicht"* discussed so far are the result of temporal sliding between the coronal, dorsal and glottal gestures in relation to the labiodental release, and due to the continuous variation along these three timing scales there is a great variability in the recorded data. But there are also instances of this word in the data base that point to a different speech production strategy. It has to do with the articulations required for the sequence **I l** being opposites: high palatal dorsum elevation with front opening and side contacts for **I**, and with front closure and side openings for **l**. The articulatory transition puts high demands on speech motor control, especially under time constraints of fast speech and unstressed syllables, and there are three possible consequences:

- **I** is adjusted to **l**, which happens in the instances of early coronal gesture timing,
- there is a short period of all-round closure (corresponding to a segment **d**), for which there are also examples in the data base,
- **l** is adjusted to **I**: the coronal gesture is eliminated: **f I l- aI C t** **[fɪ'ɛɪçt]**.

The latter process can no longer be subsumed under temporal sliding, but represents gestural reorganization: the tongue tip/blade movement is deleted from the articulatory plan, as in *"wahrscheinlich"*, discussed above.

Another few examples from spontaneous speech in the Kiel Corpus are to give illustrations of the variety of componential residues (see spectrograms in [22], pp. 14-17).

KAE g197a011 ***könnten***
canonical SAMPA **k 9 n t @ n+**
variant SAMPA **k -h ' '9 ~~ n- t-q @- n+**
IPA **[k'œ̃n̪n]**

- The first nasal consonant is deleted as a sequential element, but a residue of nasalization is still manifest in the preceding vowel as a componential feature.
- The plosive **t** is realized as glottalization somewhere in the sonorant context (vowel, nasal consonant), without a precise temporal and segmental alignment.
- In both cases the articulatory components require a non-linear symbolization, i.e. markers that do not receive durations:
  - **~~** refers to nasalization
  - **t-q** to glottalization;
  - both are aligned to the same point in time as the following, non-deleted segment **n**,
  - indexing phonetic parameters in the segmentally labelled environment (further details in [21]).

HAH g074a010 ***nicht zu spät***
canonical SAMPA **n I C t+ t s u:+ S p 'E: t**
variant SAMPA **n I C t-+ t s -MA u:-+ S p -h 'E:-'e: t**
IPA **[nɪç t̪ʷsʷʸu̯ʃʷʸp'e:t]**

- The voiced vocalic stretch of **u:** is absent;
- its lip rounding, and presumably its tongue position, remain as componential residues of labialization and velarization in the surrounding fricatives.

TIS g072a015 *kann Ihnen das*
canonical SAMPA **k a n+ Q i: n @ n+ d a s+**
variant SAMPA **k -h a n+ -MA Q- i:- n @ n-+ %d-n a s+**
IPA **[k^{h}an n^{j} nas]**

- The segment **i:** is deleted;
- its dorso-palatal tongue elevation remains as a componential residue of palatalization in the nasal consonants.

HAH g074a000 *Universitätsstädte*
canonical SAMPA **Q U n I v E6 z I t 'E: t s #S t "E t @**
variant SAMPA **%Q U n I-i: v E6 z -MA I- t -h 'E: t s-S #%S t -h "E-"e: t-d @**
IPA **[ʔʊnivɛɐ̯zt'ɛːtʃt̪^{h}ˌeːdə]**

- The segment **I** after the fricative **z** is deleted;
- its dorso-palatal elevation remains as a componential residue of palatization in **z**, which, moreover, keeps its voicing as in intervocalic position although it now occurs before **t**.

Automatic labelling, such as the output of MAUS (see the contribution by Schiel and Kipp [35]), also ought to supplement the purely linear concept of the phonetic segment by the consideration of overlapping long articulatory/acoustic components, such as glottalization, nasalization etc.. Thus if MAUS labels a stretch of speech wave as **t @ m i: l I** (as in *"das paßt mir **terminlich** schlecht"*) it most likely does not capture the actual pronunciation adequately, because the nasal feature of the deleted nasal consonant **n** presumably lingers on in the nasalization of the vowel **i:** and the sonorant **l**, and the word-final fricative **C** leaves its trace in the word-initial **S**. In such a case the Kiel labelling would insert **-~** in the first, **-MA** in the second instance. An automatic transcription has to do likewise, because it is only then that the symbolized pronunciation becomes empirically plausible; **t @ m i: l I** is not.

2.3 Content words: degrees of articulatory adjustment

The three types of interference with the phonetic unit of a word are not limited to function words. For example, in German numerals *"-zehn"* may be realized as **[tsn]**, and, over and above that, *"-zehnhundert"* (as in *"neunzehnhundert vierundneunzig"*) may even be pronounced **[tsɐt]**, as long as the word refers to a year and *"hundert"* is not stressed. In the Kiel Corpus, for instance, we find the following variant (in SAMPA notation) for *"(Mai) neunzehnhundert vierundneunzig"* **(BACg142a005);** see spectrogram in [34], pp. 162.

n 'OY n t s e:- n- #h- "U- n- d- 6 t f 'i:-'i:6 r- U- n t- #%n "OY n t s I C .
It is, on the one hand, a strongly reduced variant, linked to the citation form pronunciation
n'OYntse:n#h"Und6t f'i:rUnt#n"OYntsIC ,
on the other hand, it does not represent the end of the reduction scale because there may be further articulatory simplification, namely

- voiceless vowels in the voiceless obstruent environments
- **t** deletion before **s**
- deletion of nasal consonants and nasalization of the preceding vowels,

resulting in the variant
n 'OY -~ n- t- s e:- n- #h- "U- n- d- -MA 6- t f 'i:-'i:6 r- U- n t- #%n "OY -~ n- t- s -MA I- C.

Filling in possible further variants between the canonical form, the corpus example and the most integrated pronunciation we get the following set of IPA-transcribed word sequences from most separated to most fused:

[n'ɔɪntse:nh,ʊndɐt f'i:ʁʊntn,ɔɪntsɪç]
[n'ɔɪntsənɦ,ʊndɐt f'i:ɐnn,ɔɪntsɪç]
[n'ɔɪntsn̩ɦ,ʊnnɐt f'i:ɐnn,ɔɪntsɪç]
[n'ɔɪntsn̩,ʊnnɐt f'i:ɐnn,ɔɪntsɪç]
[n'ɔɪntsn̩ənɐt f'i:ɐnn,ɔɪntsɪç]
[n'ɔɪntsnənɐt f'i:ɐnn,ɔɪntsɪç]
[n'ɔɪntsnnɐt f'i:ɐnn,ɔɪntsɪç]
[n'ɔɪntsnɐt f'i:ɐnn,ɔɪntsɪç]
[n'ɔɪntsɐt f'i:ɐnn,ɔɪntsɪç]
[n'ɔɪnsn̥t f'i:ɐnn,ɔɪnsn̥ɪç]
[n'ɔ̃ĩsn̥t f'i:ɐnn,ɔ̃ĩsn̥ɪç]

The ordinal numbers ending in *"-zehnten"* provide further instances for the disappearance of phonetic words. Starting from canonical **[tsentən]**, the following articulatory reductions occur:
[tsentn] with ə-elision,
[tsenn̥n] with additional glottalization instead of velic elevation to signal a stop articulation,
[tsenn̥n] with breathy phonation in the nasal instead, to mark this break,
[tsenn] with the complete disappearance of the plosive reflex, which is possible in an unstressed syllable in nonfinal phrase position, e.g. before *"November"*, where we then get, for example, **[dʁ'aɪtsen nov'ɛmbɐ]**.
This means that the cardinal and ordinal numerals in *"dreizehn Novembertage"* and *"dreizehnten November"* may coalesce.

In a labial context before, e.g., *"Mal"* we may find the variants with labial assimilation **[tsempm][tsemm̩m][tsemm̥m][tsemm]**.
*"das hat er **dreizehn Mal** gemacht"* and *"das hat er zum **dreizehnten Mal** gemacht"* may then coalesce in the form **[tsem ma:l]**. The cardinal number can, however, have the further reductions **[tse ma:l]** and **[tsm ma:l]**, which seem to be impossible for the ordinal number. But the latter may be **[tsmm̩ ma:l]**.

The disappearance of an independent phonetic word and the creation of a new lexical item is also illustrated by the greeting *"n Abend"* instead of *"guten Abend"*. This extreme reduction of an adjectival form is only possible in cases of semantic "bleaching", as in this formula of phatic communion; it does not occur if the word retains its meaning, as in *"guten Appetit"*. A case from English would be *"St. Paul"* **[sm]** vs. *"a saint man"* **[seɪnt]**.

Two examples of word fusion from the Kiel Corpus of Read Speech are (in SAMPA notation):
dlms 091: *"**geben Sie mir** die Verbindung"* **g 'e: b- @- n- z i:- m i:6+**
dlms 001: *"**morgen** vormittag"* **m 'O6 -~ g- @- n-** .

3 Balance between articulatory economy and auditory distinctivity as a function of the communicative situation

The examples presented in the preceding sections suggest that word production is a compromise between articulatory economy for the speaker and acoustic distinctivity for the listener. Economy of effort in speech production is governed by a number of anatomical, physiological and temporal constraints in the speech producing apparatus that introduce directionality into reductions, such that they are not chaotic. Not just any changes, but only certain types are possible, which occur over and over again in the languages of the world and in historical sound change. For instance the development of nasal vowels is tied to the position before nasal consonants, which are in turn deleted; stops may become fricatives and approximants, and the latter may even disappear in inter-sonorant position, but the reversal of this chain is not possible.

These physically constrained tendencies to reduce effort are in their turn controlled by linguistic structures at all levels, from phonology to syntax and semantics, and therefore have different manifestations and distributions in different languages, although basic types can be generalized. Furthermore the degree of articulatory effort is governed by the precision the listener needs in order to understand, and this need is different in different speaking environments, for acoustic reasons as well as for reasons of redundancy in form and content. This redundancy is determined by the common core of linguistic context and context of situation in the widest sense between speaker and hearer, ranging from world knowledge through culture and society to the individual discourse setting.

The balance between articulatory effort and perceptual distinctivity is thus solved differently in various communication situations (cf. Lindblom's H&H theory [26]). In the lab speech situation the effect of the principle of articulatory economy is small and consequently the preservation of word identity is much greater than in read texts and even greater than in spontaneous speech taking place within delimited scenarios. This means that the study of different speaking styles [12] may be expected to yield different frequencies and different degrees of articulatory reductions or reinforcements, and are consequently a research area of great potential for gaining insight into human communication, an area that has been too much neglected for too long to the detriment of linguistic science. Modern phonetics has the theoretical and methodological tools to get on with the task and to put spoken language performance into its proper perspective vis-à-vis the linguistic imperialism of written language competence.

Because of this tug-of-war between production effort and perception ease it is also an important and interesting question how listeners manage - or why they do not manage - to decode various forms of spoken language, which may, in the case of casual spontaneous dialogue, be extremely "distorted" from the point of view of canonical word forms. The examples quoted in this paper can all be understood immediately by native speakers of German in the contexts in which they are uttered; even the strongly reduced version of *"nun wollen wir mal kucken"*, spoken by itself is quite intelligible. So listeners do not need complete phonetic signals for all the words that make up an utterance.

On the other hand, utterances that do contain all the phonetic word information may not be comprensible because they lack the necessary (non-phonetic) context of situation cues. An example is the following German sentence (in IPA transcription without word divisions and with punctuation marks to indicate sentence prosodies):
[mˈɛːənɛptəhˈɔɪ]?[nˈeː].[mˈɛːkdəmˌɛːənhˈɔɪ],[ˈɛptəbˈeːtn].

German listeners are usually not able to decode it at all - or at least not without repetition - as the pronunciation corresponding to the spelling
"Mähen Äbte Heu? Nee. Mägde mähen Heu, Äbte beten."

The hearer thus gets along with a lot less phonetic word signalling, but also needs a lot more contextual cues; how much less of the one and how much more of the other in what phonetic, linguistic and situational contexts is a question to be answered by future research.

A further, very important factor for utterance intelligibility is its prosody. What may look like a list of unconnected words on paper (Chomsk's *"furiously sleep ideas green colorless"* phenomenon), may be a perfectly structured utterance when given the right temporal, accentual and intonational properties. Jokes and English crossword puzzles thrive on this. A German example is *"Theo der Kaffee"* [tˈeːodɐkˈafeː], corresponding to the properly punctuated spellings *"Theo, der Kaffee!"* or *"Tee oder Kaffee?"*, depending on timing, intonation and pausing; but under certain prosodic conditions the utterance may stay ambiguous. This phenomenon may also be exploited across languages, as in the following example:

Un petit d'un petit	**[œ̃ptidœ̃pti]**	**Humpty Dumpty**
S'étonne au hall	**[setɔnoɔl]**	**Sat on a wall,**
Un petit d'un petit	**[œ̃ptidœ̃pti]**	**Humpty Dumpty**
Ah! degrés de folles	**[adəgʁe̯fɔl]**	**Had a great fall,**
Un dol de qui ne sort cesse	**[œ̃dɔldəkinsɔrsɛs]**	**And all the king's horses,**
Un dol de qui ne se mène	**[œ̃dɔldəkinsmɛn]**	**And all the king's men,**
Qu'importe un petit d'un petit	**[kɛ̃pɔʁtœ̃ptidœ̃pti]**	**Could not put Humpty Dumpty**
Tout Gai de Reguennes.	**[tugɛdʁəgɛn]**	**Together again.**

(Adapted from *Mots d'Heures: Gousses, Rames"*, London: Angus & Robertson (1968))

The text on the left looks French, and it also sounds French with the segmental pronunciation, transcribed in the center column, and with the appropriate French utterance prosodies added to it, but it does not make sense in French, because it is simply a string of unstructured words. But anybody familiar with the English nursery rhyme in the third column will immediately recognise it as this little poem pronounced with a heavy French accent.

The following example provides a corresponding German version of an English nursery rhyme:

Liter mies muffelt	**[lˈiːtɐmismˈʊfəlt]**	**Little Miss Muffet**
Satan atü fällt,	**[z̩ˈaːtanə̯tˈyːfɛlt]**	**Sat on a tuffet**
Hie Dinge kurz und weh.	**[hˈiːd̩ɪŋəkˈʊɐtsənvˈeː]**	**Eating her curds and whey;**
Sehr Kämme Piks beide	**[zeːɐkˈɛməpiksb̩ˈaɪdə]**	**There came a big spider,**
Ente satt Daunen bei Seide.	**[ʔɛntəz̩ˈatd̩ˈaʊnbaɪz̩ˈaɪdə]**	**And sat down beside her**
Unfrei den mies muffelt, oh weh!	**[ʔˈʊnfʁaɪdənmismˈʊfəlt ʔovˈeː]**	**And frightened Miss Muffet away.**

(Adapted from *Mörder Guss Reims"*, London: Angus & Robertson (1981))

The foregoing discussion has made it quite clear that the word may but certainly need not be a phonetic unit. The word is very flexible in its phonetic manifestation, and it can therefore not be considered "the central phonetic unit", as postulated by Tillmann's paper at this Conference.

4 Conclusion and outlook

Our knowledge about words as phonetic units in lab speech is fairly comprehensive for quite a number of languages, including English and German in particular. Phonetics has of late also been able to come to grips with the scale of decreasing word signalling from read sentences to read texts and to different types of spontaneous speech, as the data presented and interpreted in this paper testify. They show that the realisation of words by speakers is a constant interaction between phonetic integration for economy of effort on the part of the speaker, and phonetic separation for distinctivity on the part of the listener.

But this domain of phonetics above the word still requires a great deal of research, and it needs, above all, a new paradigm [19] for asking questions about pronunciation in a language. **Word phonology** has outlived itself. We have to look much more closely at the regularities of production and perception processes at the **utterance level** in actual speech communication, and this goal goes beyond the word as a phonetic unit and beyond the collection of phonetic variants lexica, because we should not just deal with the question of how the words of a language are pronounced, we also need to give answers why the pronunciations are the way they are under the constraints of the utterance in communicative context. This scientific perspective also demands a thorough integration of the symbolic domain of phonological structures with the signal domain of phonetic speech dynamics. At IPDS Kiel we have been working very intensively on the question of utterance phonology and phonetics overlapping, and interfering with, word phonology and phonetics. Our German Research Council grant will allow us to continue this work within a framework of fundamental research to gain deeper scientific insight into how speech works. The focus is on German but we are ultimately aiming at a comparative treatment of European languages [2,17,18,31,32].

5 References[1]

(AIPUK = Arbeitsberichte d. Instituts f. Phonetik u. digitale Sprachverarb. d.Univ. Kiel)

[1] DUDEN: *Das Aussprachewörterbuch.* 3. Aufl. Mannheim/Wien/Zürich: Dudenverlag (1990)

[2] Helgason, P.: Lenition in German and Icelandic. *AIPUK* 31, 219-226 (1996)

[3] Helgason, P., Kohler, K. J.: Vowel deletion in the Kiel Corpus of Spontaneous Speech. *AIPUK* 30, 115-157 (1996)

[4] IPDS: CD-ROM#1: *The Kiel Corpus of Read Speech, vol. I.* Kiel: IPDS (1994)

[5] IPDS: CD-ROM#2: *The Kiel Corpus of Spontaneous Speech, vol. I.* Kiel: IPDS (1995)

[6] IPDS: CD-ROM#3: *The Kiel Corpus of Spontaneous Speech, vol. II.* Kiel: IPDS (1996)

[1]Graphic signal representations and speech output of utterances referred to in this paper can be found at the following URL: `www.ipds.uni-kiel.de/examples.html`.

[7] IPDS: CD-ROM#4: *The Kiel Corpus of Spontaneous Speech, vol. III.* Kiel: IPDS (1997a)

[8] IPDS: *xassp* (Advanced Speech Signal Processor under the X Window System) - User's Manual. Version 1.2.15. *AIPUK* 32, 31-115 (1997)

[9] Jones, D.: *English Pronouncing Dictionary.* 15th ed. (P. Roach, J. Hartman, eds.). Cambridge: Cambridge University Press (1997)

[10] Kohler, K. J.: Glottal stops and glottalization in German. Data and theory of connected speech processes. *Phonetica* 51, 38-51 (1994)

[11] Kohler, K.J.: Complementary phonology: a theoretical frame for labelling an acoustic data base of dialogues. *Proc. ICSLP94*, vol. 1, 427-430, Yokohama (1994)

[12] Kohler, K.J.: Articulatory reduction in different speaking styles. *Proc. XIIIth ICPhS*, vol. 2, 12-19, Stockholm (1995)

[13] Kohler, K.J.: The realization of plosives in nasal/lateral environments in spontaneous speech in German. *Proc. XIIIth ICPhS,* vol. 2, 210-213, Stockholm (1995)

[14] Kohler, K.J.: Phonetic realization of German /ə/-syllables. *AIPUK* 30, 159-194 (1996)

[15] Kohler, K.J.: Phonetic realization of /ə/-syllables in German. *AIPUK* 31, 11-14 (1996)

[16] Kohler, K. J.: Articulatory reduction in German spontaneous speech. *Proc. 1st ESCA Tutorial and Research Workshop on Speech Production Modeling: from control strategies to acoustics*, 1-4, Autrans (1996)

[17] Kohler, K. J.: Glottal stop and glottalization - A prosody in European languages. *AIPUK* 30, 207-216 (1996)

[18] Kohler, K.J.: Glottalization across languages. *AIPUK* 31, 207-210 (1996)

[19] Kohler, K.J.: Developing a research paradigm for sound patterns of connected speech in the languages of the world. *AIPUK* 31, 227-233 (1996)

[20] Kohler, K. J.: Labelled data bank of spoken standard German - The Kiel Corpus of Spontaneous Speech. *Proc. ICSLP 96*, vol. 3, 1938-1941, Philadelphia (1996)

[21] Kohler, K., Pätzold, M., Simpson, A. P.: *From scenario to segment - The controlled elicitation, transcription, segmentation and labelling of spontaneous speech. AIPUK* 29 (1995)

[22] Kohler, K., Pätzold, M., Simpson, A. P.: From the acoustic data collection to a labelled speech data bank of spoken Standard German. *AIPUK* 32, 1-29 (1997)

[23] Kohler, K.J., Rehor, C.: Glottalization across word and syllable boundaries. *AIPUK* 30, 195-206 (1996)

[24] Kühnert, B.: Die alveolare-velare Assimilation bei Sprechern des Deutschen und Englischen: Kinematische und perzeptive Grundlagen. *Forschungsberichte des IPSK München* 34, 175-392 (1996)

[25] Lehiste, I.: An acoustic-phonetic study of internal open juncture. Phoentica 5 (Suppl.), 1-54 (1960)

[26] Lindblom, B.: Explaining phonetic variation: a sketch of the H &H theory. In W.J. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modelling*, 403-439. Dordrecht: Kluwer Academic Publishers (1990)

[27] Nolan, F.: The descriptive role of segments: evidence from assimilation. In D. Ladd and G. Docherty (eds.), *Papers in Laboratory Phonology II. Gesture, Segment, Prosody,* 261-280, Cambridge: Cambridge University Press (1992)

[28] Pätzold, M.: *KielDat* - Data bank utilities for the *Kiel Corpus. AIPUK* 32, 117-126 (1997)

[29] Rehor, C.: Phonetische Realisierung von Funktionswörtern im Deutschen. *AIPUK* 30, 1-113 (1996)

[30] Rehor, C., Pätzold, M.: The phonetic realization of function words in German spontaneous speech. *AIPUK* 31, 5-10 (1996)

[31] Rodgers, J.: Vowel deletion/devoicing. *AIPUK* 31, 211-218 (1996)

[32] Rodgers, J.: Vowel devoicing/deletion in English and German. *AIPUK* 32, 177-195 (1997)

[33] Rodgers, J.: A comparison of vowel devoicing/deletion phenomena in English laboratory speech and German spontaneous speech. *AIPUK* 32, 197-214 (1997)

[34] Rodgers, J., Helgason, P., Kohler, K. J.: Segment deletion in the Kiel Corpus of Spontaneous Speech. *AIPUK* 32, 127-176 (1997)

[35] Schiel, F., Kipp, A.: Probabilistic analysis of pronunciation with MAUS. Paper at ZAS *Confenerence on "The Word as a Phonetic Unit"*, Berlin Oct. 1997

[36] Simpson, A. P., Pätzold, M. (eds.): *Sound Patterns of Connected Speech - Description, Models and Explanation. AIPUK* 31 (1996)

[37] Wells, J. C.: *Pronunciation Dictionary*. London: Longman (1990)

[38] Wells, J. C., Barry, W., Fourcin, A. J.: Transcription, labelling and reference. In: A. Fourcin, G. Harland, W. Barry, V. Hazan (eds.), *Speech Technology Assessment. Towards Standards and Methods for the EUROPEAN COMMUNITY*, 141-159 (1989)

[39] WÖRTERBUCH DER DEUTSCHEN AUSSPRACHE (H. Krech et al., eds.), München: Max Hueber (1969)

Word-level phonetic variation in large speech corpora

Patricia A. Keating
Phonetics Lab, Linguistics Department, UCLA

1. Introduction

The phonetic word is of crucial importance for continuous speech recognition. This is because the word is both a basic unit that is recognized (i.e. pieces of the speech signal are matched to words in a recognition lexicon) and a basic unit used in higher-level language models. The pronunciation variability of words is very important, since such variability makes it harder to match signals to lexical items. It is particularly problematic in large vocabulary systems, since variation in the pronunciation of one word will likely make it confusable with some other word.

The problem of pronunciation variability of words has become acute as recognition has turned to more casual, unscripted, speech. Word and acoustic models built from careful speech, especially read speech, have not generalized well to more natural speech. It is thought that more natural speech is more variable in two ways:

- **phonetically** (more realizations of phonemes or other sub-word units of recognition)
- **phonemically** (more realizations of each word expressed in such units)

The typical solutions to these problems are:

- **phonetic:** use more training data to get better statistical models of acoustic variation
- **phonemic:** use more pronunciations per word in the recognition lexicon

In most automatic speech recognition systems, words are entered into a lexicon with one pronunciation ("word model") -- either from a dictionary, or some estimate of the "Most Common Pronunciation", or a baseform designed specifically as input to a phonology. Phonological rules or networks can then be used to generate alternate pronunciations from any one of these types of lexical entries. Or, alternatively, alternate pronunciations can be entered directly into a lexicon. For example, a working group at the 1996 speech recognition summer workshop reported in Fosler et al. (1996) that they tried putting pronunciations actually found in their training data (pronunciations found at least seven times) into their lexicon. There is a clear trade-off between allowing few vs. many pronunciations for each word. Cohen (1989) estimated that for careful (e.g. read) speech, a single pronunciation for each word covers (on average) about 80% of its tokens, but to cover the other 20% of tokens, multiple pronunciations are required. Thus a recognition system which performed at 59% correct using only a Most Common Pronunciation for each lexical item, improved to 66% correct under one scheme of multiple pronunciations (weighted for probability) generated by rule from a single base form. At the same time, Cohen also showed that it is crucial not to generate too many alternate pronunciations of lexical entries, else the recognizer can be overwhelmed by false alarms.
In this paper I will test the hypothesis that the pronunciation of words is more variable in unscripted speech than in read speech. If this is so, then this confounding of hits by false alarms in a lexicon with multiple pronunciations would be more problematic for unscripted

speech. If only a small number of pronunciations is allowed (because of the false-alarm problem) then many pronunciations of many words will be necessarily unrepresented in a lexicon, leading to misses. It will then become important to understand which words or word classes are likely to be more variable, so that different strategies can be applied to different parts of the lexicon.

2. Method

2.1. Speech materials

2.1.1. Corpora

The two most important large corpora of recorded American English speech are TIMIT[1] and Switchboard[2]. TIMIT consists of 6300 read sentences, 10 each from 630 speakers, totaling about 5100 word types and about 54391 word tokens. Switchboard consists of about 3 million (orthographic) word tokens of unscripted telephone conversations from 550 speakers. TIMIT was for some time the resource most used in developing and testing continuous speech recognition systems; as a result, recognizers got very good at read speech. Problems arise when everything learned from and based on TIMIT is carried over to recognizing speech from Switchboard - recognition error rates, while no longer as disastrous as they were even two years ago, are much higher.

All of TIMIT could be used for this study since it is available at little cost. A randomly chosen subset of Switchboard was available from a previous project (Keating et al. 1994).

2.1.2. Words (lexical items)

A set of words that occur in both corpora was chosen, and pronunciations of each word were compared across the corpora. For practical reasons, by "word" here is meant the orthographic word, i.e. delimited by spaces or punctuation. Thus, while "no" and "know" count as different words, "that" (determiner) and "that" (complementizer), or "like" (preposition or interjection) and "like" (verb), would count as the same word; and while "it" and "it's" would both count as single words, "it is" would count as two words. It is quite possible that some pronunciation variation of lexical items counted in this way arises from the fact that different linguistic words are being collapsed together.

To study pronunciation variability a large number of tokens is required for each word. Frequency counts for words in TIMIT (*sa*, *sx*, and *si* sentences) were made from our database (REF). Frequency counts for the 160 most common words in Switchboard, and frequency bins for about 100 other words of variable frequency in Switchboard, were made available by Mark Liberman (p.c.). An arbitrary threshold for inclusion was set at 33 tokens per word, that is, a word must occur at least 33 tokens in each corpus. A further criterion, which applied only to the TIMIT sample, was that no more than 3 of the tokens for a word could come from the same speaker or the same (orthographic) sentence. No attempt was made to eliminate tokens that occurred within identical word strings shorter than the sentence. (This means that in Switchboard, more than in TIMIT, some tokens may have come from similar contexts. So this would work to reduce the apparent variability in Switchboard, and thus make Switchboard and TIMIT more alike in variability (thus going against the hypothesis)).

---

[1] DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) October 1990, NIST Speech Disc 1-1.1 (1 disc); http://www.nist.gov/itl/div894/894.01/corpora/timit.htm

[2] http://www.ldc.upenn.edu/ldc/catalog/html/speech_html/scr.html

Table 1. Comparison sample: words sampled from both Switchboard (SWB) and TIMIT, in alphabetical order, with total number of tokens of each word in each corpus. There are about 3 million word tokens in Switchboard, about 54000 in TIMIT. Therefore to compare the two figures very approximately, multiply the TIMIT figure by 50. To compare Switchboard with Kučera and Francis (1967), a 1 million-word corpus, divide the Switchboard figure by 3.

WORD	*# tokens in SWB*	*# tokens in TIMIT*
a	72924	1168
about	12362	50
and	106833	667
are	14024	349
as	10141	197
at	10791	134
be	14321	263
but	28291	136
don't	18641	668
for	19867	377
had	11033	709
have	30394	149
he	9594	341
I	121443	127
in	40532	1260
is	26182	517
it	55571	236
like	23441	697
my	15007	117
not	14977	158
of	56340	640
on	17010	267
one	12728	78
or	16851	117
out	11091	82
so	26417	65
that	67035	827
the	98301	2202
them	10468	58
there	13290	59
they	33212	179
this	9862	210
to	73147	1370
up	9973	89
was	24187	321
we	25672	187
well	22024	37
what	14933	62
with	14044	244
you	80241	362

A total of 40 of the 60 most common words in Switchboard, words that also occur in TIMIT, were selected by these criteria and are listed in Table 1. (High-frequency words of Switchboard that occur infrequently or not at all in TIMIT are: *uh, yeah, uh-huh, that's, think, oh, really, right, um, I'm*, and words which occur fewer than 33 times, or which failed the second criterion, are: *know, it's, don't.*) To insist on more than 33 tokens greatly limits the number of words that can be studied in TIMIT, and of course those that do occur this often are all high-frequency function words. So as to include some lower-frequency words, including content words, in the study, an additional 32 words were selected from Switchboard only. These are shown in Table 2. The pronunciations of these words cannot be compared to TIMIT, but they can be compared to the high-frequency function words in Switchboard.

Table 2. Sample of other words from Switchboard only (not enough tokens occur in TIMIT). Exact frequencies not available, only frequency ranges.

WORD	*frequency in SWB*
after	between 1000 and 1400
cases	between 40 and 50
chips	between 40 and 50
could	between 3000 and 5500
down	between 3000 and 5500
facts	between 40 and 50
glass	between 180 and 240
goal	between 40 and 50
island	between 100 and 140
know	47560 (included here because too few in TIMIT)
market	between 180 and 240
metric	between 100 and 140
must	between 300 and 500
okay	between 3000 and 5500
once	between 1000 and 1400
paint	between 180 and 240
played	between 300 and 500
probably	between 3000 and 5500
road	between 180 and 240
simple	between 100 and 140
since	between 1000 and 1400
stick	between 180 and 240
system	between 1000 and 1400
taken	between 300 and 500
there's	between 3000 and 5500
under	between 300 and 500
upon	between 100 and 140
very	between 3000 and 5500
weeds	between 40 and 50
weekend	between 300 and 500
what's	between 1000 and 1400
years	between 3000 and 5500

2.1.3. Tokens

Matched numbers of tokens of each word were selected at random from Switchboard and TIMIT. This number was determined by whichever corpus yielded the smaller number of tokens (usually TIMIT). The number of tokens from each corpus was capped at 40.

2.2. Transcriptions

Phonetic and phonemic (dictionary-style) transcriptions of each token were obtained. Throughout this paper these transcriptions are shown in the ARPAbet-style symbols of the TIMITbet (Zue and Seneff 1988), listed in Table 3.

For TIMIT, the phonetic transcriptions used were those provided with the corpus: the "TIMITbet" transcriptions which are narrower than phonemic, but not especially narrow. Phonemic transcriptions were derived from these by a set of collapsing rules which collapsed the phonetic categories into fewer, broader, categories. The general approach of the collapsing rules is to map each more-specific symbol into the phonetically most similar more-general symbol. These collapsing rules do not take into account what the word is.

Table 3. TIMITbet symbols and nearest IPA equivalents; phonemic symbols used here. Case is not distinctive for TIMITbet symbols

TIMITbet symbols	*nearest IPA symbol*	*phonemicized here as*
pcl	p˺ (closure only)	p
p	p (release only)	p
b	b (release only)	b
bcl	b˺ (closure only)	b
t	t (release only)	t
tcl	t˺ (closure only)	t
d	d (release only)	d
dcl	d˺ (closure only)	d
k	k (release only)	k
kcl	k˺ (closure only)	k
g	g (release only)	g
gcl	g˺ (closure only)	g
f	f	f
v	v	v
th	θ	th
dh	ð	dh
s	s	s
z	z	z
sh	ʃ	sh
zh	ʒ	zh
ch	tʃ (release only)	ch
jh	dʒ (release only)	jh

h (or hh)	h	h
hv	ɦ	h
m	m	m
n	n	n
ng	ŋ	ng
em	m̩	ax m
en	n̩	ax n
eng	ŋ̩	ax ng
r	ɹ	r
l	l	l
er (also listed below)	ɻ	ax r
el	l̩	ax l
w	w	w
y	j	y
dx	ɾ	d
nx	ɾ̃	n
q	ʔ	-
iy	i	iy
ih	ɪ	ih
ey	eɪ	ey
eh	ɛ	eh
ae	æ	ae
aa	ɑ	aa
ay	aɪ	ay
aw	aʊ	aw
ao	ɔ	ao
ow	oʊ	ow
oy	ɔɪ	oy
uh	ʊ	uh
uw	u	uw
ah	ʌ	ah
er	ɝ	ax r
ux	ʉ	uw
ix	ɨ	ih
ax	ə	ax
ax-h	ə̥	ax
axr	ɚ	ax r

For Switchboard, the initial phonetic transcriptions were done at UCLA and were narrower still, in "UCLAbet" symbols (Keating et al. 1994). Some of the Switchboard transcriptions were done by two or more transcribers. Agreement between these transcribers was good overall for unscripted telephone speech. Therefore additional Switchboard transcriptions were done by the author alone. It should be noted that in general it seems harder to get transcribers to agree when transcribing rapid fluent speech like Switchboard, than when

transcribing read speech like TIMIT. Thus, pronunciation variability is probably necessarily confounded with transcription variability in studies such as the one here (with human transcribers).

These narrow transcriptions have been done for the purpose of studying phonetic variation in more detail than TIMITbet transcription would allow. For present purposes, however, these were converted into TIMITbet by a second set of collapsing rules. The Switchboard phonemic transcriptions were then derived from these TIMITbet transcriptions as was done for TIMIT. Table 4 schematizes the levels of transcription.

Table 4. Levels of transcription produced by collapsing rules.

	UCLAbet narrow	--->	*TIMITbet phonetic*	--->	*phonemic*
TIMIT	(not available)		yyy	--->	zzz
SWB	xxx	--->	yyy	--->	zzz

Another difference between the corpora relevant to the transcriptions is that while Switchboard is telephone speech, TIMIT is not (at least, not the original TIMIT used for the transcriptions). So to the extent that Switchboard is degraded speech relative to TIMIT, that could also make the pronunciations seem more variable -- it is simply harder to ascertain what the speaker said. In fact though this is probably not a big factor here: when a sample was really noisy we didn't use it, and the difficult issues of transcription were not generally related to bandwidth or noise. (They were about syllabicity and vowel reduction.)

2.3. Analyses

From the set of transcriptions, the Most Common Pronunciation was determined for each word in each corpus at each level of transcription. The *Most Common Pronunciation*, or MCP, is that pronunciation that occurs most frequently in the sample of 33-40 tokens, and its *coverage* is the percentage of the sample with that pronunciation. For example, 39 of 40 tokens of "stick" have the phonemic transcription /s t ɪ k/, so that is its MCP (phonemic), and the coverage of that MCP is 98%.

A number of different counts and calculations were also done. These will be described along with their results in sections below.

3. Results

3.1. Number of pronunciations per word

The raw number of distinct pronunciations was counted for each word. These are summarized in Table 5 for the 40 words available for both corpora.

Table 5. Average numbers of pronunciations per word, comparison sample of 40 words.

	in TIMIT	*in SWB*
phonetic transcriptions	9.5	14.3
phonemic transcriptions	5.8	9.5

It can be seen that there are fewer different phonemic pronunciations than phonetic in both corpora (this is almost definitionally so), and that there are fewer different pronunciations at both levels in TIMIT than in Switchboard, as hypothesized. These results can be compared with those in Table 6, which shows the same counts for the sample of 32 other words from Switchboard, mostly low-frequency content words. The figures for these words in Switchboard are remarkably similar to those for the higher-frequency words in TIMIT.

Table 6. Average numbers of pronunciations per word, lower-frequency words (SWB only)

phonetic transcriptions	10.0
phonemic transcriptions	5.7

3.2. Phonemic variation

It is quite striking that even in a phonemic (dictionary-style) transcription, there are almost 10 different pronunciations per high-frequency word for samples of only 33-40 words, and over 5 different pronunciations even for lower-frequency words. Phonemic transcriptions were tabulated because it is sometimes suggested that if only the phonemes could be reliably recovered from the signal, then the word recognition problems would be minor. The results in the previous section show that this is not true. (In a similar vein, Fosler et al. (1996) compared (hand-done) Switchboard transcriptions with dictionary baseforms, and found that on average, one out of eight phones (phonemes) from the baseforms were deleted in the transcriptions.) However, the figures in Tables 5-6 are averages, and it is certainly the case that some words do not vary much in phonemic transcriptions. For those words, which are listed in Table 7, successful recognition of the phonemes would ensure ready recognition of the words. While such words are generally from the low-frequency sample, it can be seen that not all 32 low-frequency words have this property, as there are only 10 such words here.

Table 7. Words in Switchboard (out of 72) which do not vary much at phonemic level.

WORD	*# phonemic pronunciations*
bear	2
facts	2
glass	2
goal	2
like	2
metric	3
must	3
my	3
simple	3
stick	2
system	2
very	2

For those words which do vary at the phonemic level, several generalizations can be made, which hold for the content words too. All phonemic pronunciations which occurred four or more times were examined and the following patterns found.

3.2.1. The 2-schwas problem: TIMITbet distinguishes between a lower [ax] and a higher [ix] reduced vowel (basically, IPA [ə] vs. [ɨ]). The criterion for deciding between them is whether F2 is closer to F1 vs. F3. These two reduced vowels were phonemicized differently, as /ax/ vs. /ih/. In general this accords with the underlying vowels, but not always. For some words individual tokens were found to vary in the F2 frequency, and this difference was then carried up to the phonemicization. Note that these phonemicizations are determined only by the signal; it would be circular to restore underlying segments on the basis of lexical knowledge. Words with this variation included *a, and, as, at, but, cases, in, is, of, system, taken, that, the, was, what, with.*

3.2.2. Vowel reductions: In general, all vowels in function words can reduce. There were some general tendencies in these reductions, as follows (in IPA symbols): /i/ /u/ /ʊ/ often reduce to /ɪ/; /ʌ/ /o/ /ɛ/ often reduce to /ə/; /æ/ often reduces to /ɛ/. But there was enough variation beyond these patterns to give rise to multiple pronunciations, in words such as *and, as, be, but, could, don't, one, she, so, that, them, under, we, what, what's, you.*

3.2.3. Flapping: Both underlying /t/ and /d/ were often flapped. However, all flaps were phonemicized as /d/, since that is the phonetically closer quality.

3.2.4. Final /t d n l/ loss: These anterior coronal consonants tend to not be heard/seen word-finally, but not consistently so. Words with this variation included *and, at, don't, down, in, it, must, not, out, paint, road, that, weekend, well, what.*

3.2.5. Dialect variation in vowels: Some words contain vowels that seem to vary greatly across speakers, including *my, on, our, the, well, I.*

3.2.6. Weak syllable loss: Stressless syllables are vulnerable in vowel-initial iambs (*upon, about*) and word-medially (*probably*), but not consistently so.

3.2.7. Initial /dh/ loss in function words: Words like *them, they, this* may appear to lose their initial consonant in some, but not all, contexts. They are particulary vulnerable when following another function word ending in a nasal.

3.2.8. Final -(r)z devoicing: Word-final /z/ is sometimes devoiced in *there's, years.*

Table 8. Phonemic MCP and its coverage in the two corpora; dictionary pronunciation (converted to phonemic transcription used here). In the dictionary consulted, some special r-colored vowel symbols were used; these have been converted here to our usual transcriptions. Where the MCP for a given word is different in the two corpora, the coverage of each MCP in the other corpus is given in parentheses.

WORD	*MCP in SWB*	*its coverage*	*MCP in TIMIT*	*its coverage*	*dictionary form*
a	ax	29	ax	55	ey, ax
about	ax b aw /	8	ax b aw t	39	ax b aw t
	ax b aw t /		(ax ba w	6)	
	b aa / ih b aw d		(b aa	0)	
[illegible]	eh n	25	ih n	35	ae n d, ax n

	(ih n	15)	(eh n	28)	ax n d, en
are	ax r	62	aa r	54	aa r, ax r, axr
	(aa r	30)	(ax r	30)	ax
as	ih z	53	ih z	32	ae z, ax z
at	ih t	24	ae t	29	ae t, ax t
	(ae t	5)	(ih t	20)	
be	b iy	67	b iy	100	b iy, b ih
but	b ah t	26	b ah t	38	b ah t, b ax t
don't	d ow n	33	d ow n t	45	d ow n t
	(d ow n t	0)	(d ow n	39)	
for	f ax r	61	f ax r	61	f ao r, f ax r
had	h ae d	51	h ae d	38	h ae d
have	h ae v	69	h ae v	67	h ae v
he	h iy	66	h iy	87	h iy
I	ay	64	ay	92	ay
in	ih n	45	ih n	79	ih n
is	ih z	71	ih z	87	ih z
it	ih t	36	ih t	62	ih t
(know)	n ow	86	n ow	91	n ow
like	l ay k	97	l ay k	100	l ay k
my	m ay	71	m ay	87	m ay
not	n aa t	43	n aa t	55	n ao t
of	ax v	32	ax v	49	ah v, ao v, ax v
on	ao n	32	ao n	41	ao n
one	w ah n	50	w ah n	84	w ah n
or	ax r	64	ao r	41	ao r, ax r
	(ao r	5)	(ax r	31)	
out	aw / aw t	20	aw t	54	aw t
			(aw	14)	
so	s ow	61	s ow	87	s ow
that	dh ih t	15	dh ae t	23	dh ae t, dh ax t
the	dh ax	35	dh ax	40	dh iy, dh ax, dh ih
them	ax m /	26	dh eh m	72	dh eh m, dh ax m
	dh ax m /		(ax m	3)	
	dh eh m		(dh ax m	8)	
there	dh eh r	69	dh eh r	36	dh eh r
they	dh ey	68	dh ey	95	dh ey
this	dh ih s	69	dh ih s	89	dh ih s
to	t ih	31	t uw	31	t uw, t ax
	(t uw	23)	(t ih	28)	
up	ah p	59	ah p	95	ah p
was	w ih z	39	w ih z	33	w ah z, w ao z, w ax z
we	w iy	66	w iy	89	w iy
well	w eh l	34	w eh l	86	w eh l
what	w ax d/w ax t	15	w ah d	35	w ao t, w ah t
	(w ah d	13)	(w ax d	5)	
			(w ax t	8)	
with	w ih th	41	w ih th	54	w ih th, w ih dh
you	y uw	35	y uw	78	y uw

3.3. Most Common Pronunciation

Recall that Cohen (1989) found that the MCP covers, on average, about 80% of tokens for words in read speech. Table 8 gives the phonemic MCP, and its coverage, for each word in our comparison samples. It also gives a pronunciation for each word taken from a dictionary (Harcourt, Brace, & World's *Standard College Dictionary*, 1963). Table 9 gives the phonetic MCPs and their coverage. For this sample, the MCP is often the same for the two corpora. Phonemically, it is the same for 80% of the words, while phonetically it is the same for 65% of the words. That is, a phonemic lexicon based on the MCPs in TIMIT is a reasonable starting point for a Switchboard lexicon, since the agreement here is 80%. Furthermore, when the MCP's coverage is greater than 50% in both corpora (that is, just the cases where the MCP is doing the most work), the two corpora almost always have the same MCP. Exceptions to this generalization are phonetic *this* (TIMIT [dh ih s], Switchboard [dh ix s]) and phonemic *are* (TIMIT /aa r/, Switchboard /ax r/).

Table 9. Phonetic MCP and its coverage in the two corpora. Format as in previous table.

WORD	*MCP in SWB*	*its coverage*	*MCP in TIMIT*	*its coverage*
a	ix	26	ax	47
	(ax	24)	(ix	21)
about	ax bcl b aw q	8	ax bcl b aw tcl	25
	(ax bcl b aw tcl	3)	(ax bcl b aw q	3)
and	eh nx / en	13	ix n	18
	(ix n	8)	(eh nx	5)
			(en	8)
are	axr	41	aa r	38
	(aa r	11)	(ax r	19)
as	ix z	47	ix z	23
at	ix tcl	15	ae tcl	15
	(ae tcl	2)	(ix tcl	10)
be	bcl b iy	49	bcl b iy	59
but	bcl b ah dx	15	b ah tcl	23
	(b ah tcl	0)	(bcl b ah dx	0)
don't	dcl d ow n	15	dcl d ow n tcl	24
	(dcl d ow n tcl	0)	(dcl d ow n	6)
			(dcl d aw nx	5)
for	f ax r	51	f ax r	56
had	hv ae dx	16	eh dcl/hv ae dx	16
	(eh dcl	0)		
have	hv ae v	38	hv ae v	38
he	hv iy	37	hh iy	79
	(hh iy	21)	(hv iy	8)
I	ay	32	ay	54
in	ih n / ix n	16	ix n	39
			(ih n	21)
is	ix z	42	ix z	53
it	ih q	12	ih tcl	24
	(ih tcl	0)	(ih q	2)
(know)	n ow	57	n ow	91
like	l ay kcl k	49	l ay kcl k	59

my	m ay	61	m ay	87
not	n aa tcl	33	n aa tcl	53
of	ax	30	ax v	41
	(ax v	27)	(ax	5)
on	ao n	24	ao n	35
one	w ah n	25	w ah n	70
or	axr	38	axr	26
out	aw / aw tcl	11	aw tcl	46
			(aw	0)
so	s ow	61	s ow	87
that	dh ae dx	13	dh ae tcl	18
	(dh ae tcl	5)	(dh ae dx	5)
the	dh ax	30	dh ax	35
them	dh eh m	26	dh eh m	72
there	dh eh r	67	dh eh r	36
they	dh ey	65	dh ey	95
this	dh ix s	60	dh ih s	86
	(dh ih s	9)	(dh ix s	3)
to	t ix / tcl t ix / tcl t ux	13	tcl t ix	21
			(t ix	8)
			(tcl t ux	10)
up	ah pcl p	38	ah pcl p	49
was	w ix z	39	w ax z	31
	(w ax z	25)	(w ix z	28)
we	w iy	63	w iy	89
well	w eh l	31	w eh l	86
what	w ax dx	15	w ah dx / w ah tcl	33
	(w ah dx	13)	(w ax dx	5)
	(w ah tcl	8)		
with	w ix th	31	w ix th	44
you	y ix	20	y ux	63
	(y ux	20)	(y ix	10)

It can readily be seen also that for most words the MCP has better coverage in TIMIT than in Switchboard: this is so for 78% of the words considered phonemically, and 80% of the words considered phonetically. There are some exceptions, however; the words *are, as, for, have, had, or, there, was* are more consistently reduced in Switchboard, so that the MCP is this reduced form.

The average coverages are given in Table 10. At both levels of transcription there is about a 15% difference in coverage. That means that, although a TIMIT-based lexicon in general will provide a good base form for Switchboard, the coverage offered by that form will be less. It will be noted that these coverages are quite low in general; in particular, the 62% phonemic coverage in TIMIT is much lower than Cohen's 80% figure for read speech. This is in part because the sample here is limited to a set of very high-frequency function words, whereas Cohen's figure was derived over a larger set of words. In addition, Cohen's data were not from TIMIT, but from a study of the DARPA Resource Management Database[3], which involves only a subset of the speakers from TIMIT, reading database query sentences.

[3] http://www.itl.nist.gov/div894/894.01/corpshrt.htm

Table 10. Coverage of MCP (% of sample) -- comparison sample of 40 words		
	in TIMIT	*in SWB*
phonetic	48	33
phonemic	62	47

Table 11 shows that the average coverage of the phonemic MCP for the lower-frequency words in Switchboard is 70%, much closer to Cohen's 80%. These low-frequency words in Switchboard are more like the high-frequency words in TIMIT above. So we would expect a lexicon derived from TIMIT to work reasonably for the lower-frequency content words of Switchboard, but not for the high-frequency function words. These two tables also show that the difference between the two samples from Switchboard (higher frequency words in Table 10, lower frequency words in Table 11) is greater when phonemic transcriptions are counted.

Table 11. Coverage of MCP (% of sample) -- Switchboard-only sample of 32 words	
phonetic	47
phonemic	70

The phonemic MCP can be compared to a dictionary entry, shown in the last column of Table 8. The dictionary consulted here included alternate reduced pronunciations for function words. In general these pronunciations correspond to the observed MCP (plus some British-like variants given in the dictionary): they are the same for 90% of the 40 words for TIMIT, and for 75% of the (same) 40 words for Switchboard.

Finally, it is interesting to see whether any words within these samples share their MCP, or look as if they might share their MCP with some other word not in the sample. Such cases would pose obvious problems for recognition. There are a few, whether the phonetic or the phonemic transcriptions are considered. In the TIMIT sample, *and/in*, *as/is*, and *are/our* share their MCP, and in Switchboard *as/is* and *are/or* do (see tables for specific forms).

3.4. Other schemes for inclusion of pronunciations

3.4.1. Pronunciations occurring 7 or more times

Fosler et al. (1996) attempted to improve recognition performance by constructing a recognition lexicon from observed pronunciations. Pronunciations observed at least 7 times in the training data, a sample of 2116 sentences, were used. What kind of coverage would this criterion give for the present Switchboard samples? While the number of word tokens in the 2116 sentences that they sampled is larger than the number of tokens in the present study, the number of high-frequency words is probably roughly similar. For the samples here of 33-40 tokens per word, a pronunciation that occurs 7 times would cover about 18-21% of the tokens.

Table 12 gives the number of pronunciations occurring 7 or more times for each word. It can be seen that there are usually 1 or 2 per word; the average is 1.45 such pronunciations per word. For those words where there is one such pronunciation, or two which are tied in coverage, it is the same as the MCP. For other words with 2 such pronunciations, their combined coverage will necessarily be better than that of the MCP pronunciation alone. But for a few words, there is no such pronunciation - no single pronunciation occurs at least 7 times - and for these words, this criterion would hurt, not help, coverage.

Table 12. Counts of phonemic transcriptions, high-frequency Switchboard sample only.

WORD	*# prons 7+ times*	*coverage (%)*	*# prons 2+ times*	*coverage (%)*	*#prons 50% coverage*
a	2	55	7	92	2
about	0	0	10	67	5
and	2	48	6	85	3
are	2	92	2	92	1
as	1	48	7	93	2
at	1	25	9	90	3
be	1	67	3	87	1
but	2	44	7	92	3
don't	1	33	3	61	3
for	1	63	4	93	1
had	1	51	5	89	1
have	1	69	3	82	1
he	1	66	5	92	1
I	2	83	3	93	1
in	1	45	4	79	2
is	1	71	5	97	1
it	2	60	5	88	2
like	1	97	1	97	1
my	2	92	3	100	1
not	2	75	4	95	2
of	3	89	4	95	2
on	2	57	5	92	2
one	2	75	3	83	1
or	1	64	5	90	1
out	2	40	7	74	4
so	2	82	4	95	1
that	0	0	8	83	4
the	2	60	5	90	2
them	3	77	5	90	2
there	1	69	4	86	1
they	1	68	4	100	1
this	1	69	3	89	1
to	2	54	8	92	2
up	1	59	5	85	1
was	2	58	5	89	2
we	2	92	2	92	1
well	1	34	6	86	2
what	0	0	7	78	4
with	1	41	6	82	2
you	2	58	5	85	2

3.4.2. Pronunciations occurring more than once

Table 12 also shows the number of pronunciations per word when a less restrictive criterion is applied: eliminate only pronunciations that occur only once (the presumed outlier pronunciations). These pronunciations cover, on average, 88% of the tokens for the 40

words in the Switchboard sample (with an average of five pronunciations per word), and virtually the same coverage, 89%, for the second Switchboard sample (with an average of three pronunciations per word). The coverage of these pronunciations ranges from 61% to 100%, but is generally high. Still, this result means that the outlier pronunciations which have been excluded account for over 10% of the tokens. Furthermore, this figure of 3-5 pronunciations per word, which seems to be necessary to get even this moderately acceptable level of covereage, is a high number, when the false alarm problem of high-vocabulary recognition is considered.

3.4.3. Pronunciations giving 50% coverage of samples

Finally, Table 12 shows the number of pronunciations per word needed for 50% coverage. Here we see numbers of pronunciations per word that would not cause a large false alarm problem.

4. Conclusion

This study has compared pronunciation variability, for a set of 40 lexical items, in the read speech of TIMIT vs. the non-read speech of Switchboard. The read speech of TIMIT is less variable on every measure. The Most Common Pronunciations of the lexical items are often the same in the two corpora (the same for 80% of lexical items sampled, in phonemic transcription), but their coverage is much reduced -- only 57% of the individual tokens for the 72 words in the two Switchboard samples presented here. More pronunciations beyond this one Most Common Pronunciation must be allowed to get reasonable coverage of the tokens of at least the high-frequency lexical items. Even if phonemes can be recognized completely accurately, there will still be much pronunciation variability to deal with.

The results presented here show that this variability is not the same for all words, however. The low-frequency content words of Switchboard vary no more than do words in TIMIT; therefore these words should present no new difficulties. It is the high-frequency function words of Switchboard that vary so much more and which must be the focus of new efforts. Even with these, not all of them vary greatly, or if they do, not always in ways that would make them potentially confusable with other lexical items. Therefore it would seem that research should focus on strategies for just the most variable and confusable words. For example, since these are function words, perhaps better language models for the structures they occur in could help.

Another possible approach would be to focus more on phonetic variation that distinguishes one sequence of phonemes from another. Even the TIMITbet-style transcriptions studied here collapse over phonetic details that could be useful in distinguishing lexical items, details that can be spread out over a larger span of speech. Some of these details are well-known to phoneticians: vowel nasalization that distinguishes *and* from *at*, or *in* from *it*, even when final consonants are deleted; glottalization that also distinguishes *in* from *it*; vowel duration differences that preserve voicing distinctions, or reflect the number of underlying consonants in a word. Other differences are less well-known, being either idiosyncratic or prosodic: for example, that *our* and *are* are generally distinguished by nasalization in *our*; or by the presence of a full glottal stop at the beginning of *our*. Such differences as these are not reflected in the transcriptions compared in this study. Furthermore, for many such useful properties, there are currently no good acoustic models that would allow their recognition. I

hope phoneticians and phonologists will get to work on this challenge, which provides a chance to show that our knowledge of sound structure can help with a practical problem.

References

Cohen, M. H. (1989). *Phonological structures for speech recognition*. Ph.D. dissertation, Dept. Electrical Enginerring and Computer Science, UC Berkeley.

Fosler, E., M. Weintraub, S. Wegmann, Y.-H. Kao, S. Khudanpur, C. Galles, M. Saraclar (1996). Automatic Learning of Word Pronunciation From Data. *ICSLP-96*.

Keating, P., M. MacEachern, A. Shryock, and S. Dominguez (1994). A manual for phonetic transcription: Segmentation and labeling of words in spontaneous speech. Manual written for the Linguistic Data Consortium, *UCLA Working Papers in Phonetics 88*, 91-120.

Kučera, H. and N. Francis (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Zue, V. and S. Seneff (1988), Transcription and alignment of the TIMIT database, *Proc. Second Meeting on Advanced Man-Machine Interface through Spoken Language*, pp. 11.1-11.10.

Acknowledgments

This work was supported by the UCLA Academic Senate Committee on Research. I thank Chai-Shune Hsu and Narineh Hacopian for much-appreciated assistance.

PROBABILISTIC ANALYSIS OF PRONUNCIATION WITH 'MAUS'

Florian Schiel, Andreas Kipp

Institut für Phonetik und Sprachliche Kommunikation,
Ludwig-Maximilians-Universität München

ABSTRACT

This paper describes a method to automatically detect pronunciation variants in large speech corpora within the framework of the 'MAUS' project ([1]). 'MAUS' stands for 'Munich Automatic Segmentation System' and is a general purpose tool to automatically label and segment read and spontaneous German speech into phonetic/phonologic segments. The output of MAUS can for example be used to build probabilistic models of pronunciation of fluent German as reflected by the analysed corpus. These models can be the basis for phonetic investigations or can be incorporated into classic speech recognition algorithms.

The paper is organised as follows: The first section gives a very short introduction into the main processing principle of MAUS and gives some examples of the output of MAUS applied to utterances from the Verbmobil corpus. Section 2 deals very briefly with the problem of how to evaluate such an output. A method is given that first compares the performance of three human transcribers with each other and then the performance of MAUS with each of them. Section 3 describes our method for deriving probabilistic pronunciation dictionaries from the MAUS output and gives some interesting examples from the Verbmobil domain. The 4th and last section gives some new approaches towards incorporating these models into a new automatic speech recognition (ASR) approach that combines phonetically 'sharper' acoustic models with the probabilistic modelling of pronunciation.

1. INTRODUCTION TO MAUS

The MAUS system was developed at the Bavarian Archive for Speech Signals (BAS) to facilitate the otherwise very time-consuming manual labeling and segmentation of speech corpora into phonetic units. Initially funded by the German government within the Verbmobil I project, MAUS is now further extended by BAS with the aim to automatically improve all BAS speech corpora by means of complete broad phonetic transcriptions and segmentations. The basic motivation for MAUS is the hypothesis that automatic speech recognition (ASR) of conversational speech as well as high quality 'concept-to-speech' systems will require huge amounts of carefully labelled and segmented speech data for their successful progress.

Traditionally a small part of a speech corpus is transcribed and segmented by hand to yield bootstrap data for ASR or basic units for concatenative speech synthesis (e.g. PSOLA). Examples for such corpora are the PhonDat I and II corpus (read speech) and the Verbmobil corpus (spontaneous speech). However, since these labelings and segmentations are done manually, the required time is about 800 times the duration of the utterance itself, e.g. to label and segment an utterance of 10 sec length a skilled phonetician spends about 2 h and 13 min at the computer. It is clear that with such an enormous effort it is impossible to annotate large corpora like the Verbmobil corpus with over 33 h of speech. On the other hand large databases are needed urgently for empirical investigations on the phonological and lexical level.

Input to the MAUS system is the digitised speech wave and any kind of orthographic representation that reflects the chain of words in the utterance. Optionally there might be markers for non-speech events as well, but this is not essential for MAUS. The output of MAUS is a sequence of phonetic/phonemic symbols from the extended German SAM Phonetic Alphabet ([5]) together with the time position within the corresponding speech signal.

Example:

```
Input:
Speech Wave + 'bis morgen wiederhoeren'

Output:
MAU: 0 479 -1 <p:>
MAU: 480 480 0 b
MAU: 961 478 0 I
MAU: 1440 1758 0 s
MAU: 2720 959 1 m
MAU: 3680 799 1 O
MAU: 4480 2399 1 6
MAU: 6880 2079 1 N
MAU: 8960 799 2 v
MAU: 9760 959 2 i:
MAU: 10720 479 2 d
MAU: 11200 2239 2 6
MAU: 13440 799 2 h
MAU: 14240 639 2 2:
MAU: 14880 1439 2 6
MAU: 16320 1599 2 n
MAU: 17920 1759 -1 <p:>
```

(The output is written as a tier in the new BAS Partitur format. 'MAU:' is a label to identify the MAUS tier; the first integer gives the start of the segment in samples counted from the beginning of the utterance; the second integer gives the length of the segment in samples; the third number gives the word order and the final string is the labeling of the segment in extended German SAM-PA. See [10] for a detailed description of the BAS Partitur format)

MAUS is a three-staged system (see fig.1):

In a first step the orthographic string of the utterance is looked up in a canonical pronunciation dictionary (e.g. PHONOLEX, see [8]) and processed into a Markov chain (represented as a directed acyclic graph) containing all possible alternative pronunciations using either a set of data driven microrules or using the phonetic expert system PHONRUL.

A microrule set describes possible alterations of the canonical pronunciation within the context of $+/-1$ segments together with the probability of such a variant. The microrules are automatically derived from manually segmented parts of the corpus. Hence, these rules are corpus dependent and contain no a priori knowledge about German pronunciation. Depending on the pruning factor (very sel-

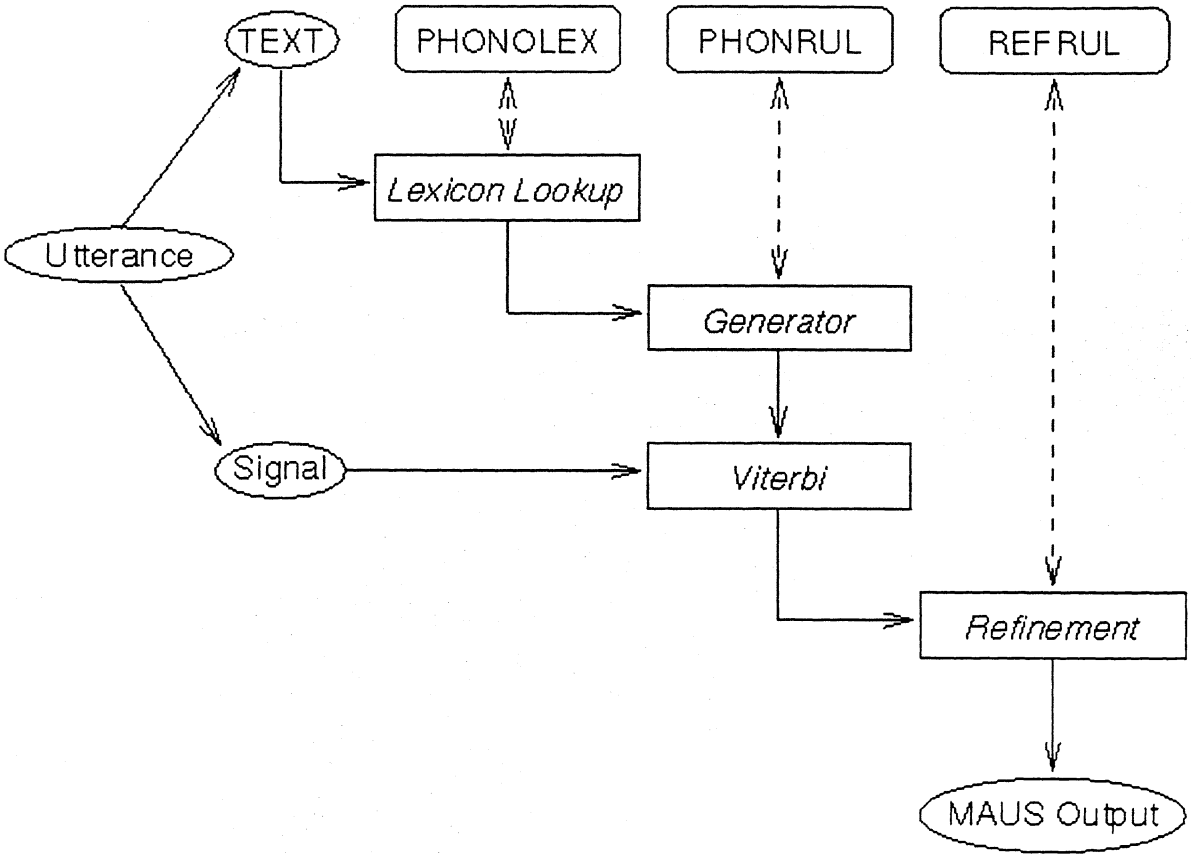


Figure 1: The MAUS system - block diagram

dom observations are discarded) and the size of the manually segmented data the microrule set consists of 500 to 2000 rules. In this paper we use a set of approx. 1200 rules derived from 72 manually segmented Verbmobil dialogs of The Kiel Corpus of spontaneous Speech ([6]). Details about this method can be found in [1].

The expert system PHONRUL consists of a rule set of over 6000 rules with unlimited context. The rules were compiled by an experienced phonetician on the basis of literature and generalised observations in manually transcribed data. There is no statistical information within this rule set; all rules are treated with equal probability. PHONRUL is therefore a generic model and should be considered independent of the analysed speech corpus. A more detailed description of PHONRUL can be found in [7].

The second stage of MAUS is a standard HMM Viterbi alignment where the search space is constrained by the directed acyclic graph from the first stage (see figure 2 for an example). Currently we use the HTK 2.0 as the aligner ([9]) with the following preprocessing: 12 MFCCs + log Energy, Delta, Delta-delta every 10 msec. Models are left-to-right, 3 to 5 states and 5 mixtures per state. No tying of parameters was applied to keep the model as sharp as possible. The models were trained to manually segmented speech only (no em-

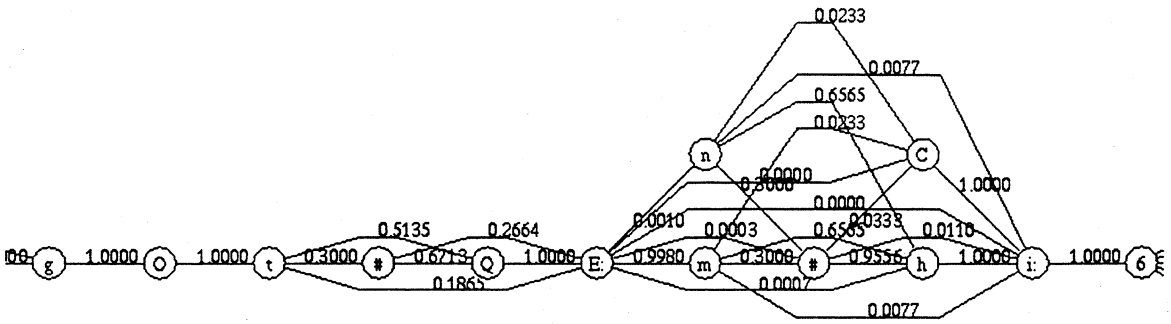


Figure 2: Acyclic graph of the utterance "Gott... ähm... hier..." with possible pronunciation variants

bedded re-estimation).

The outcome of the alignment is a transcript and a segmentation of 10 msec accuracy, which is quite broad. Therefore in a third stage REFINE the segmentation is refined by a rule-based system working on the speech wave as well as on other fine-grained features. However, the third stage cannot alter the transcript itself, only the individual segment boundaries.

The general drawback of the MAUS approach is, of course, that MAUS cannot detect variants that are not 'foreseen' by the first stage of the process. However, we found that using the microrule method the vast number of distinct rules are found after analyzing a relatively small subportion of the whole corpus. This indicates that the number of non-canonical pronunciations occurring in a certain domain such as the Verbmobil corpus is in fact limited and therefore treatable by a limited number of rules.

2. EVALUATION

The output of MAUS can be separated into two different classes: the transcript (the chain of symbols) and the corresponding segmental information (begin and end of each segment).

Unlike in an ASR task the evaluation of a phonetic/phonemic segmentation of arbitrary utterances has a great disadvantage: there is no reference. Even very experienced phoneticians will not produce the same segmentation, not even the same transcript on the same speech wave.

We tried to circumvent this general problem by first comparing the results of three experienced human transcribers on the same corpus with each other to get a feeling for what is possible and set an upper limit for MAUS. We used standard Dynamic Programming techniques as used in ASR evaluations (e.g. [9]) to calculate the inter-labeller agreement between different transcripts. We found that the coverage of the three human transcribers ranges from 78.8% to 82.6% (on the basis of approx. 5000 segments). We then calculated the accuracy for the MAUS output with regard to each set of human results and found values ranging from 74.9% to 80.3% using the microrule method and 72.5% to 77.2% using PHONRUL. Not surprisingly, the worst and best coverage were correlated in all three experiments. This means that if we set the upper limit to the best match within human tran-

scription results (82.6%) and compare this to the average agreement of MAUS with these two human transcribers, we'll end up with a relative performance of 97.2% for MAUS. (Note that this relative performance measure might be higher than 100% at some distant point in the future!)

For a more detailed discussion about the problem of evaluation as well as a more accurate analysis of the MAUS output (applied to read speech) please refer to [3].

In terms of accuracy of segment boundaries the comparison between manual segmentations shows a high agreement: on average 93% of all corresponding segment boundaries deviate less than 20msec from each other. The average percentage of corresponding segment boundaries in a MAUS versus a manual segmentation is only 84%. This yields a relative performance of 90.3%. We hope that a further improvement of the third stage of MAUS will increase these already encouraging results.

3. PROBABILISTIC PRONUNCIATION MODEL

Aside from the many other uses of the MAUS output for this paper we'll show how to derive a simple but effective probabilistic pronunciation model for ASR from the data. There are two obvious ways to use the MAUS results for this purpose:

- use direct statistics of the observed variants
- use generalised statistics in form of microrules

In the following we will discuss both approaches.

3.1. Direct Statistics

Since in the MAUS output each segment is assigned to a word reference level (Partitur Format, see [10]), it is quite easy to derive all observed pronunciation variants from a corpus and collect them in a PHONOLEX ([8]) style dictionary. The analysis of the training set of the 1996 Verbmobil evaluation (volumes 1-5,7,12) led to a collection of approx. 230.000 observations.

The following shows a random excerpt of the resulting dictionary:

```
terminlich
adj
t E 6 m i: n l I C
t E 6 m i: n I C        3
t @ m i: l I C  3
t E 6 m i: n l I C      10
t E 6 m i: l I C        1
t @ m i: n l I C        7
&
...
Karfreitag
nou
k a: 6 f r aI t a: k
k a: 6 f r aI t a: k    15
k a: 6 f r aI t a x     3
&
...
weil
par
v aI l
v a l   11
v aI    108
v aI l  207
&
...
siebenundzwanzigsten
adj
z i: b @ n U n t t s v a n t s I C s t @ n
z i: b @ n U n s v a n t s I s t @ n    1
z i: b m U n s v a n t s I k s t n      2
z i: b m U n s v a n s I C s t n        1
z i: b m U n s v a n t s I C s t @ n    1
z i: m U n s v a n t s s t @ n  1
z i: m U n s v a n t s s n      1
z i: b m U n s v a n t s s t    1
s i: b @ n U n s v a n s I C s t n      1
```

```
z i: b @ n U n s v a n s I C s t n     1
z i: b @ n U n s v a n t s I s t n     3
z i: m U n s v a n t s I s t @ n       1
z i: b m U n s v a n s I z n   2
i: b m U n s v a n z I z n     1
z i: m U n s v a n t s I z n   6
z i: m U n s v a n t s I s n   1
z i: b m U n s v a s I s t n   1
z i: b @ n U n s v a n t s I C s t n   2
z i: m U n s v a n s I s n     1
z i: m U n s v a n z I k s t @ n       1
z i: b m U n s v a n t s I z n 2
z i: m U n s v a n t s I s t   2
z i: b m U n s v a n t s I s t @ n     2
z i: m U n s v a n s I s t n   17
z i: m U n s v a n s s t n     1
z i: b m U n s v a n s I C s t  1
z i: m U n s v a n t s I C s n 1
z i: b m U t s v a n t s s t n 1
z i: m U n s v a n s I k s t @ n       2
z i: b m U n s v a n s I s t n 6
z i: b m U n s v a n s I k s t n       1
z i: b m U n s v a n s I s t @ n       4
z i: m U n s v a n t s I k s t n       6
z i: m U n s v a n t s s t n   1
z i: m U n s v a n s I z n     1
z i: b m U n s v a n s s t n   2
z i: b m U t s v a n s I s t n 1
z i: m U n s v a s s n  1
z i: b @ n U n s v a n t s I k s t n   2
z i: m U n s v a n t s I C s t n       9
z i: m U n s v a n z I z n     2
z i: m U n s v a n t s I s t n 27
z i: b m U n s v a n t s s t n 1
z i: m U n s v a n z I s t n   5
z i: b m U n s v a n s I s n   1
z i: m U n s v a n s s n       1
z i: m U n s v a n s s t @ n   1
z i: b m U n s v a n z I z n   1
z i: m U n s v a n t s I C s t @ n     2
s i: b m U n s v a n t s I C s t n     1
z i: b m U n s v a n t s I C s t n     12
z i: m U n s v a n s s t       1
z i: m U n s v a n s I C s t n 3
z i: m U n s v a n z I s n     1
z i: b m U n s v a n t s I s t n       28
z i: m U n s v a n s I s t @ n 3
z i: b m U n s v a n t s I k s t @ n   1
&
...
Namen
nou
n a: m @ n
n a: m  30
n a: m @ n      15
&
...
Essen
nou
Q E s @ n
@ s n   2
E s n   16
E s @ n 6
s n     3
E s     1
Q E s @ n       7
Q E s   1
Q E s n 21
&
```

The above modified PHONOLEX format is defined as follows:

```
<orthography>
<comma separated list of
linguistic classes>
<canonical pronunciation>
<empiric pronunciation> <count>
&
...
```

Obviously many of the observations are not frequent enough for a statistical parameterisation. Therefore we prune the baseline dictionary in the following way:

- Observations with a total count of less than N per lexical item are discarded.
- From the remaining observations for each lexical word L the a-posteriori probabilities $P(V|L)$ that the variant V was observed are calculated. All variants that have less than M% of the total probability mass are discarded.
- The remaining variants are renormalised to a total probability mass of 1.0.

Applied to the above example this yields the following more compact statistics (pruning parameters: N=20, M=10):

```
terminlich              0.434783
t E 6 m i: n l I C
terminlich              0.130435
t E 6 m i: n I C
terminlich              0.304348
t @ m i: n l I C
terminlich              0.130435
t @ m i: l I
Karfreitag              1.000000
k a: 6 f r aI t a: k
weil                    0.342857
v aI
weil                    0.657143
v aI l
siebenundzwanzigsten    0.509091
z i: b m U n s v a n t s I s t n
siebenundzwanzigsten    0.490909
z i: m U n s v a n t s I s t n
Namen                   0.333333
n a: m @ n
Namen                   0.666667
n a: m
Essen                   0.320000
E s n
Essen                   0.420000
Q E s n
Essen                   0.120000
E s @ n
Essen                   0.140000
Q E s @ n
```

where the second column contains the a-posteriori probabilities. This form can be directly used in a standard ASR system with multi pronunciation dictionary like HTK (version 2.1).

3.2. Generalised statistics

The usage of direct statistics has the disadvantage that most of the words will be modelled by only one variant, which in many cases will be the canonical pronunciation because of lack of data. An easy way to generalise to less frequent words (or unseen words) is to use not the statistics of the variants itself but the underlying rules that were applied during the segmentation process of MAUS. Note that this has nothing to do with the statistical weights of the microrules mentioned earlier in this paper; it's the number of appliances of these rules that counts. Since there is formally no distinction between microrules for segmentation in MAUS and probabilistic rules for recognition, we can use the same format and formalism for this approach as in MAUS. The step-by-step procedure is as follows:

A: Derive a set of statistical microrules from a subset of manually segmented data or use the rule set PHONRUL (see section 1).

B: Apply this rule set to segment the training corpus and count all appliances of each rule forming the statistics of the recognition rule set.

Note that the recognition rule set might be a subset of the PHONRUL/microrule set, although this is very unlikely for the latter.

This approach has the great advantage that the statistics are more compact (and therefore robust), independent of the dictionary used for recognition (which for sure will contain words that were never seen in the training set) and generalise knowledge about pronunciation to unseen cases. However, the last point may be a source of uncertainty, since it cannot be foreseen whether the generalisation is valid to all cases where the context matches. We cannot be sure that the context we are using is sufficient to justify the usage of a certain rule in all places where

this context occurs.

4. AUTOMATIC SPEECH RECOGNITION (ASR)

There have been several attempts to incorporate knowledge about pronunciation into standard methods for ASR. Most of them (with some exceptions, e.g. [4]) didn't yield any improvements. The argument was that the advantage of a better modelling on the lexical level is eaten up by the fact that the search space and/or the dictionary ambivalence increases. However, most of the literature did not take into account reliable statistics (because they were simply not available) and used acoustic models that were trained using canonical pronunciations. Our hypothesis is that an increase in recognition performance can only be achieved if the following two conditions are satisfied:

1. A reliable statistical model for pronunciation (which very likely will be adapted to the task).
2. Acoustical models that match the modelling on the lexical level.

On this basis we are currently conducting several experiments with a standard HTK recogniser for the 1996 Verbmobil evaluation task. In this paper we will only report about preliminary results using the direct statistics approach of section 3.1.

A standard recogniser of HTK 2.0 with the following properties was designed for the experiment:

The speech signal is mean subtracted, emphasised and preprocessed into 12 MFCCs + log Energy, Delta, Delta-delta every 10 msec. Training and test sets are defined in the 1996 Verbmobil evaluation task ('Kuer', test corpus: 6555 words). The canonical dictionary contains 840 different entries. The language model is a simple bigram calculated exclusively from the training set. The acoustic models are monophone left-to-right HMMs with 3-5 states containing a variable number of mixtures without tying. We use 46 models from the extended German SAM-PA including one model for silence and one model for non-speech events.

We trained and tested the recogniser with the same amount of data in two different fashions:

- *Baseline System*
  Standard bootstrapping to manually labelled data (1h40) and iterative embedded re-estimation (segmental-k-means) using 30h of speech until the performance on the independent test set converged (note: performance in terms of word accuracy, defined by (number of words - insertions - replacements - deletions) / number of words). The re-estimation process used a canonical pronunciation dictionary with one pronunciation per lexical entry.
  The system was tested with the same canonical dictionary.

- *MAUS System*
  This system was bootstrapped to one third of the training corpus (approx. 10h of speech) using the MAUS segmentation and then iteratively re-estimated (30h of speech) using not the canonical dictionary but the transcripts of the MAUS analysis (note that the segmental information of the MAUS analysis is NOT used for the re-estimation).

The system was tested with the probabilistic pronunciation model described in section III.1. using the pruning parameters N=20 and M=0%.

Figure 3 shows the performance of both systems during the training process. Note that the MAUS system starts with a much higher performance because it was bootstrapped to 10h of MAUS data (compared to 1h40min of manually labelled data for the baseline system). After training, the MAUS system converges on a significantly higher performance level of 66.35% compared to 63.44% of the baseline system.

5. CONCLUSION

The MAUS system can be used effectively to fully automatically label and segment read and spontaneous speech corpora into broad phonetic alphabets. This enables us for the first time to derive statistical models on different processing levels (acoustic, phonetic, lexical) on the basis of very large databases. We have shown that the usage of this data can significantly improve ASR on spontaneous speech.

The MAUS principle is not language dependent (however, the required resources are!). Therefore we strongly encourage colleagues in other European countries to adopt the MAUS principle for their specific languages and produce similar resources as are currently produced at BAS for the German language. A first joint project (MIGHTY MAUS) for American English and Japanese is scheduled for 1998 together with the International Computer Science Institute (ICSI), Berkeley California, and Sofia University, Tokyo.

REFERENCES

[1] A. Kipp. M.-B. Wesenick, F. Schiel (1997): Pronunciations Modelling Applied to Automatic Segmentation of Spontaneous Speech; in Proceedings of the EUROSPEECH 1997 Rhodes, Greece, pp. 1023-1026.

[2] A. Kipp, M.-B. Wesenick, F. Schiel (1996): Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 106-109.

[3] A. Kipp, M.-B. Wesenick (1996): Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 129-132.

[4] T. Sloboda, A. Waibel (1996): Dictionary Learning for Spontaneous Speech; in: Proceedings of the ICSLP Philadelphia, pp. 2328.

[5] SAM Phonetic Alphabet, www.phon.ucl.ac.uk/home/sampa/home.htm

[6] IPDS (ed.): The Kiel Corpus of Spontaneous Speech; CDROM 1 + 2, University of Kiel, 1995.

[7] M.-B. Wesenick (1996): Automatic Generation of German Pronunciation Variants; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 125-128, Oct 1996.

[8] PHONOLEX, www.phonetik.uni-muenchen.de/Bas/BasPHONOLEXeng.html
or
F. Schiel (1997): The Bavarian Archive for Speech Signals; in: FIPKM 1997, Institute of Phonetics, University of Munich, to appear.

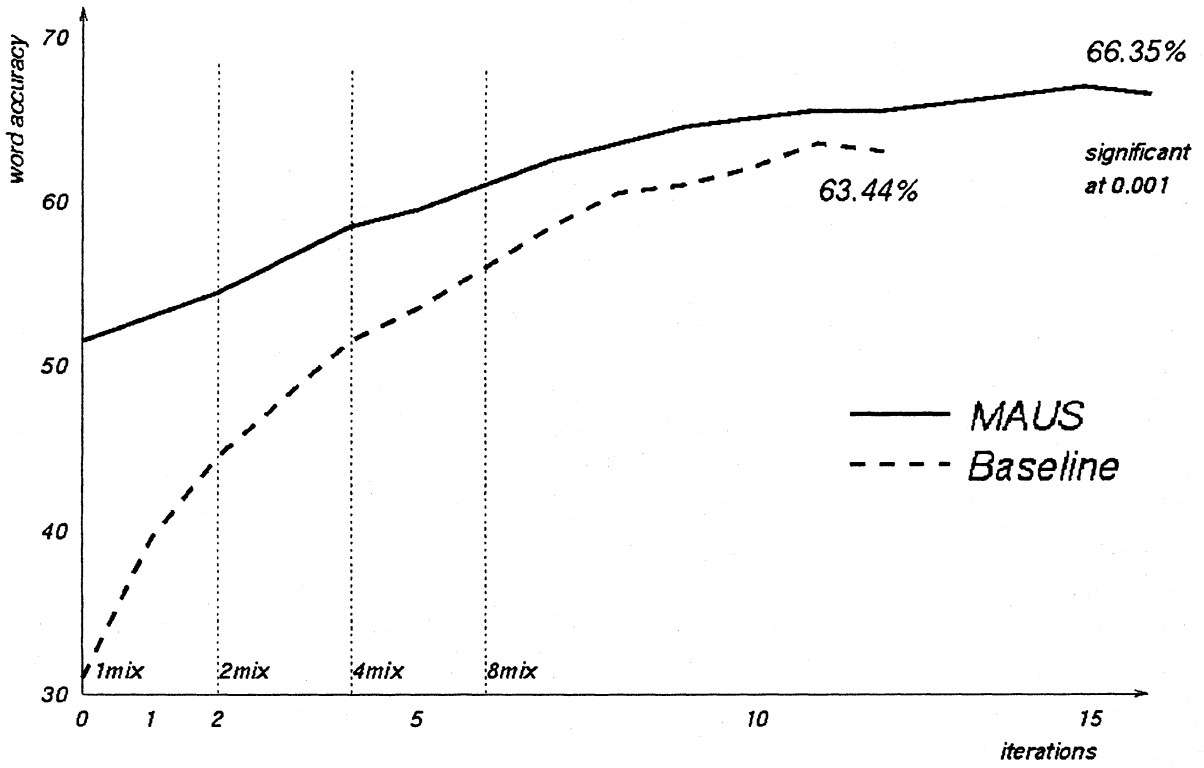


Figure 3: Performance of baseline system compared to the system trained with MAUS data and probabilistic pronunciation model

[9] S. Young et al. (1995/1996): The HTK Book; Cambridge University.

[10] Partitur Format, www.phonetik.uni-muenchen.de/Bas/BasFormatsdeu.html
or
F. Schiel, S. Burger, A. Geumann, K. Weilhammer (1997): The Partitur Format at BAS. In: FIPKM 97, Institute of Phonetics, University of Munich, to appear.

Introduction to the PhonDat Database of Spoken German

Christoph Draxler
(abridged version of Practical Applications of Prolog '95 conference paper)

Department of Phonetics and Speech Communication
Ludwig-Maximilians-University Munich
Schellingstr. 3
D - 80799 Munich
Tel: +49 +89 2866 9968
draxler@phonetik.uni-muenchen.de[1]

Abstract

The PhonDat project within the German Verbmobil research initiative aims at creating and making accessible a very large database of symbolic and signal data of spoken high German. Currently, the PhonDat database consists of one corpus of sentences containing all phoneme combinations of high German, and of one corpus of sentences from a train enquiries scenario. All symbolic data is held in a Prolog system with a powerful database management system extension; signal data is stored in external files.

The database is accessed through queries over the symbolic data. The result of a query evaluation is either again symbolic data, or a reference to signal files and signal fragments within these files. Two access modes are supported: a toolbox of predefined high-level query predicates for standard, albeit complex, queries; and the full Prolog programming language for custom applications. The PhonDat database interacts with external signal analysis and display applications through interprocess communication.

The PhonDat database has been used in various research applications, e.g. speech recognition training, segmentation comparisons, and statistical phoneme analyses. It is now being extended to hold the Verbmobil multi-language spontaneous speech corpus collected in a scheduling scenario.

Keywords: Phonetic database, PhonDat, Prolog, spoken language processing

1. Introduction

Spoken language processing (*SLP*) deals with the relationship between speech signal, symbolic transcriptions of signals, and orthographic text. SLP is considered one of the key technologies in telecommunications. Speech data collection is actively pursued in many countries, and SLP applications are now becoming available on workstations and PCs.

In SLP, the term *database* refers to a body of speech material. Such a database consists of signal data of recorded speech, a symbolic representation of the signal data, and administrative data. Examples of such databases are TIMIT [19], Verbmobil/PhonDat [14], KTH [1], and

1. This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil project under grant 413-4001-01IV 102 L/4.

SAM [15]; they are used as reference databases in research and for the development of SLP applications, e.g. speech recognition, speech synthesis, speech verification, speaker identification, etc.

Currently, SLP databases consist of little more than the data itself: access to and manipulation of the data has to be programmed explicitly in any application using such a database. Only recently have there been attempts to store data independent of applications in relational (*RDB*), object-oriented (*OODB*), or deductive database systems (*DDB*).

Early attempts based on RDBs (e.g. [8], [11], [12], [16]) have proven to be too restricted for SLP data. The relational model does not support structured or bit-stream data; futhermore, recursion is needed to access internal elements of structures and for the computation of the transitive closure. OODBs provide complex datastructures including bit-stream data. In general, the data manipulation language is integrated into a programming language for full computational power, e.g. CLOS in [10], C++ in [2]. However, there is a considerable semantic gap between the logic based formalisms used in phonology and linguistics and OODB languages. Furthermore, there is not yet a standard OODB language. DDBs provide complex datastructures and a powerful logic based data manipulation language; however, bit-stream data is difficult to handle in DDBs.

For the implementation of the PhonDat/Verbmobil database a hybrid approach using a persistent Prolog system was chosen for the following reasons:

- Prolog is very close to first order predicate logic, the predominant formalism in linguistics.
- Powerful and efficient Prolog environments with access to external or internal persistent databases are available.
- The implementation of database query languages on top of Prolog is straightforward.
- External applications can be accessed through foreign language interfaces and interprocess communications.

Finally, one of the main tasks in SLP is that of mapping signal, transcription, and orthography. This is a complex alignment problem involving multiple representation levels, which can be described elegantly in a logic based formalism and implemented efficiently in Prolog.

The structure of the paper is as follows: in chapter 2 the PhonDat database is presented together with the terminology needed in later chapters. Chapter 3 gives an overview over applications for which the PhonDat database has been used already and chapter 4 discusses performance, query language and interprocess communication issues. Chapter 5 gives a short conclusion.

2. The PhonDat database

The PhonDat database consists of two corpora: PhonDat I is made up of 200 phonemically balanced artificial German sentences plus the North wind fable and a Butter story [9]. PhonDat II consists of 64 pseudo spontaneous (i.e. dialogues recorded at a train enquiries counter were transcribed orthographically and read in a studio) speech sentences from a train enquiries scenario [14].

2.1 Symbolic data

At the International Phonetics Asscociation (*IPA*) meeting in Kiel in 1989 a convention for the computer representation of individual languages (*CRIL*) was defined. According to CRIL, SLP data should be represented on three symbolic levels: orthographic representation, phonemic representation, and phonetic transcription. PhonDat strictly adheres to the CRIL convention.

The *orthographic* representation is a standard text representation. In PhonDat, a *sentence* is a sequence of words. A *word* is a sequence of characters; words are delimited by blanks or punctuation marks.

```
Guten Tag, wann geht morgen vormittag ein Zug nach Frankfurt?
```

In the *citation form* (or *phonemic*) representation a reference pronunciation of words (spoken in isolation in high German) is given. In PhonDat, an extended SAM-PA alphabet is used [14], [15]. Each word in the dictionary has a citation form or *canonic word* associated to it. A canonic word is a sequence of canonic units. A *canonic unit* is a sequence of SAM-PA characters.

```
Feiertag            f 'aI6 t a: k
Frankfurt           f r 'a N k f U6 t
Freitag             f r 'aI t a: k
frueh               f r 'y:
fuer                f y:6+
geben               g 'e: b @ n
...
```

The *phonetic transcription* of an utterance is the result of a segmentation and labelling procedure. In PhonDat, the segmentation and labelling is performed on a phonetic workbench [20] by trained human segmenters or an automatic segmentation program [17]; the output of a segmentation session is a *transcription*.

A *transcription* is associated to a signal, a segmenter, and a segmentation session. It consists of a sequence of segments. A *segment* has a begin time and a duration, and a *label* which is either a marker symbol for paralinguistic information or an action operator with canonic units as arguments.

```
13031 #c:
13031 ##%g
13735 $'u:-'u
14846 $t
15924 $@-
15924 $n
18641 ##t
19437 $'a:
22085 $k
24905 #,
24905 #p:
...
```

2.2 Signal data

Signal data is stored in signal files which consist of a header for administrative data (date of recording, speaker, sampling rate. recording setup, etc.) and the signal data. In PhonDat, data is recorded at a sampling rate of 16 KHz with 16 bit encoding (i.e. 256 KB/s). Typically, a signal file is about 80 to 300 KB of size for an utterance.

2.3 Size of the database

Segmentation and labelling of the PhonDat I database is still under way – only a fraction of the longer texts have been transcribed, whereas the PhonDat II database is now completed. The current sizes of the databases are given in (Tab. 1)

	PhonDat I	PhonDat II
Speakers	200	15
Signal files	> 20000	> 3000
Size (signal files)	4 GB	600 MB
Transcriptions	> 3400	> 5200
Dictionary entries	> 350	> 200

Tab. 1: PhonDat I and PhonDat II database sizes

2.4 Prolog representation of the PhonDat database

The PhonDat database is held in six Prolog relations

```
ipa(IPA,IPAName,Type).
is_a(SuperType,SubType).
sampa_ipa(Sampa,IPA,SampleWord).
word_canword(WId,Word,CanonicWord,FunctionWord).
word_in_sentence(WId,Corpus,SentNo,Pos).
segment_file(FileName,Speaker,City,Segmenter,SentNo,Version,Segments).
```

`ipa/3` matches IPA symbol codes to IPA symbol names and articulatory properties:

```
?- ipa(308, 'lower-case u', Type).
Type = [back, high, rounded]
```

IPA symbols may be classified according to their phonetic properties, and a *type hierarchy* is constructed by successive abstraction from these properties (the top-most element in the hierarchy is a *phone*; on the next lower level there are *vowel*, *consonant* and *diacritic symbol*, etc.). The third argument of `ipa/3` contains the type information, and the hierarchy itself is stored in the `is_a/2` relation (cf.[10] for a similar type hierarchy).

`sampa_ipa/3` maps SAM-PA symbols to IPA symbols and gives a reference word for each SAM-PA symbol.

```
?- sampa_ipa(u,IPA,'Kulisse').
IPA = 308
```

`word_canonic_word/4` is the dictionary of the PhonDat database. `WId` is a system provided unique integer identifier for a word (to distinguish homographs), and `FunctionWord` is a marker.

```
?- word_canword(WId, 'Zug', CanonicWord, FunctionWord).
CanonicWord = [[103], [132], [501, 308, 503], [109]]
FunctionWord = nf
WId = 201
```

word_in_sentence/4 captures the occurrence of a word within a sentence

```
?- word_in_sentence(201, Corpus, SentNo, Pos).

Corpus = train
SentNo = 501
Pos = 9
```

segment_file/7 contains all data relevant to a phonetic transcription. Segments is a list of tuples segment(Begin,Duration,ErrorMeasure,Label), where Label is either a paralinguistic label (e.g. sentence_begin, word_begin, punctuation(Code), error), or a phonetic label. A phonetic label is either a tuple elision(CanUnit), insertion(CanUnit), replacement(CanUnit1, CanUnit2), a pause label, or a canonic unit.

```
?- segment_file(FileName,'AWE','D','CHK',501,0,Segments).

FileName = 'AWED5010.S1'
Segments = [segment(12849, 0, 0, sentence_begin),
segment(12849, 0, 0, word_begin),
segment(12849, 886, 0, arbitrary([110])),
segment(13735, 1111, 0, [501, 308, 503]),
... ,
segment(68838, 0, 0, punctuation(977))]).
```

2.5 Implementation

The PhonDat database is implemented in Eclipse, the logic programming environment of ECRC [4], and LPA MacProlog on the Macintosh.

The user selects the preferred signal display and/or analysis applications and queries the symbolic database. The result of a query in the symbolic database is either output to the screen, a reference to one or more signal fragments within signal files, or is added to the symbolic database as new knowledge.

The signal analysis and display applications are external applications. The PhonDat database accesses them via remote procedure calls or message passing (via AppleEvents on the Macintosh and TCP/IP under UNIX). The external applications receive from the database addresses within signals and can then load the corresponding signal fragment from the signal database (e.g. on CD-ROM).

3. PhonDat Application Programs

The PhonDat application programs can be divided into three categories

- Conversion programs
- Predefined query predicates
- Complex applications

All PhonDat database predefined queries and sampel application programs are described in [3].

3.1 Conversion programs

Conversion programs convert a given form into another representation. Typical examples are the translation of SAM-PA to IPA and back, or the translation of user input into an internal format.

sampa_ipa/2 converts canonic words, canonic units and segments into the corresponding IPA representation. It is implemented to work in both directions.

user_input/2 requires in its first argument a type specification in a Prolog string (or atom) and returns the corresponding IPA representation.

```
?- user_input(''a:,plosive,f',IPA).

IPA = [[501,324,503],plosive,[103]]
```

3.2 Predefined query predicates

Predefined query predicates constitute the toolbox for the phonetic database – they provide the basic set of operations which are of interest to the phonetician.

As an example, the predicate type_in_canonic_unit(Type,CanUnit) checks whether the canonic unit CanUnit is of a given type Type.

```
?- type_in_canonic_unit([*,vowel],[501,308]).

yes
```

The following complex goal checks whether the canonic word corresponding to a word contains canonic units of the type vowel, and if so, returns the SAM-PA form of the typed canonic units.

```
?- word_canonic_word(_,Word,CanWord,_), member(CanUnit,CanWord),
     type_in_canonic_unit([*,vowel,*], CanUnit),
     sampa_ipa(SampaCU,CanUnit),
     sampa_ipa(SampaCW,CanWord).

Word = 'Aachen'
CanWord = [[113],[501,304,503],[140],[322],[116]]
CanUnit = [304]
SampaCU = 'a:
SampaCW = Q 'a: x @ n
```

The set of predefined query predicates consists of approx. 40 predicate definitions.

3.3 Complex applications

Predefined query predicates are ideal for ad-hoc queries to the database. For statistical computations which compute values over collections of solutions, a more powerful database access is needed. Furthermore, the efficiency of a query evaluation of a predefined query predicate depends on the instantiation of its arguments. Application programs, for which the instantiation of arguments does not change, can be optimized for efficiency.

Currently, Prolog application programs have been developed for these tasks:

- Labelling consistency comparisons
- Analysis of vowel duration in high German

Labelling consistency comparisons

Transcriptions of the same signal produced by *i*) different and *ii*) the same segmenters were compared to allow a classification of phones into clear and unclear cases with respect to the transcription labels used [5], and with respect to the segment boundaries [6]. The main result is that the intraindividual consistency is not significantly better than interindividual consistency and that the consistency depends strongly on the class of phone found in the signal.

Analysis of vowel duration

For a complete phonetic theory of a language, duration data for the phone inventory of the langauge is needed. The PhonDat I corpus contains all dyadic phoneme combinations of high German, and is thus well suited for the analysis of durations. Furthermore, since the speakers were instructed to speak in a casual speaking style, the recorded speech is close to natural speech.

The main objective of these analyses is to gather empirical data for phonetic theories for spoken German.

4. Discussion

Two aspects of the implementation are now being discussed.

- Performance
- Interapplication communication between the database and external applications

4.1 Performance

The main goal of the PhonDat database implementation has been to provide the functionality needed for the tasks to be worked on, and efficiency has been a lesser objective only. However, efficiency is fully sufficient for interactive work.

The following benchmarks (Eclipse 3.4.5 on a Sparc 10) may serve to give an impression of the performance. The database contains 207 dictionary entries, 677 word occurrences, and 5268 segmentations with a total of over 350.000 segments:

- Compute the number, sum, and the average duration of all phonetic (i.e. non-zero length) segments: 68.5 s
- For each type class of phonemes, if there exists more than one manual segmentation of an occurrence of a canonic word, compare the manual segmentations with the automatic segmentation: 236.6 s

4.2 Interapplication communication

The PhonDat database is one tool among others for workers in the field of SLP. As such, communication with other applications is of great importance since this allows to make use of the many signal analysis, signal display, and other signal processing applications.

On the Macintosh, AppleEvents are used to communicate with other applications. For example, to display the signal corresponding to a given type, the database is queried and the resulting signal address is sent to a display software together with a command to select the appropriate signal fragment, to compute its spectrogram, and to output it via speakers (Fig. 1):

In Eclipse, the interprocess communication built-ins are used to drive external applications via sockets in the TCP/IP domain. This includes network access to the PhonDat database, as well as interchange with applications on remote machines.

5. Conclusion

Despite its restricted amount of transcription data, the PhonDat database has shown to be a valuable tool for empirical studies in Phonetics. The expressive power of the database queries is very high, and the class hierarchy of articulatory descriptors is a powerful means of formulating queries.

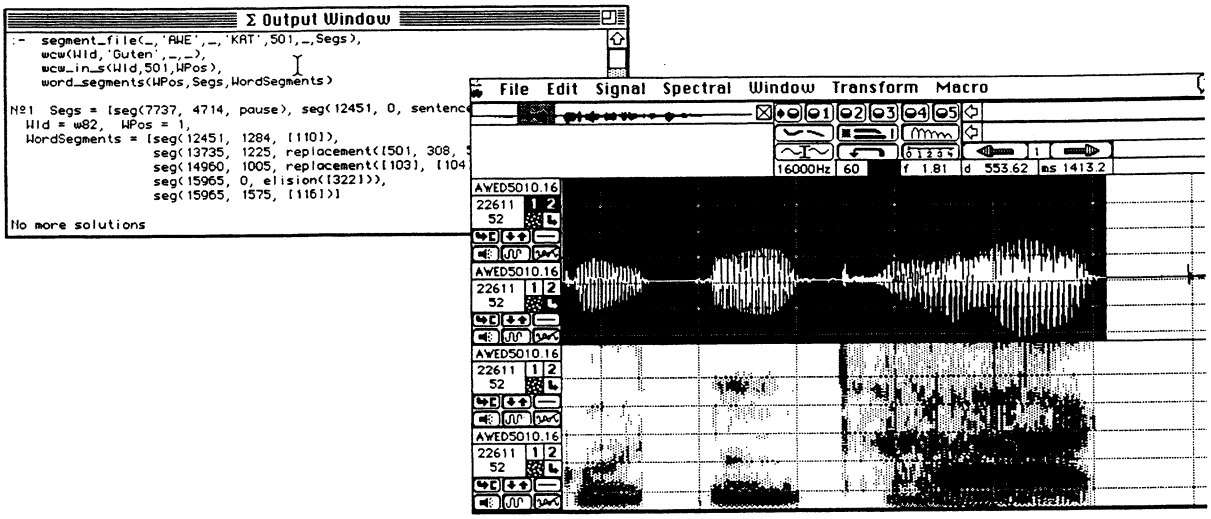


Fig. 1: Signal Output for Database Query

However, Prolog is not well suited as a database query language, at least for novice users: identifying arguments by their position instead of by name is problematic for predicates with more than five arguments. Efficiency has never been a problem; the advantage of having portable code clearly outweighs any platform specific optimizations.

Future work will include predicates that call sound processing applications into the database language, e.g. for filtering, signal computations, etc. Finally, additional layers of symbolic information, e.g. prosody, should be integrated into the database.

References

[1] R. Carlson, B. Granström, L. Nord: The KTH Speech Database. Speech Communication Vol. 9, No. 4, 1990

[2] SPEX: S. Swagten: Speech Processing EXpertise centre, Leidschendam, The Netherlands: private communication

[3] C. Draxler: The PhonDat Database Handbook. Internal report, Institut für Phonetik, Universität München, 1994

[4] ECL^{i}PS^{e}: ECRC Common Logic Programming System, User Manual, ECRC 1992

[5] B. Eisen, H. Tillmann, C. Draxler: Consistency of Judgements in manual labelling of phonetic segments: The distinction between clear and unclear cases. ICSLP Banff, 1992

[6] B. Eisen: Reliability of Speech Segmentation and Labelling at Different Levels of Transcription, Eurospeech 93, Berlin

[7] J. Esling: Computer Coding of the IPA: Supplementary Report. Journal of the International Phonetic Association, vol 20, No 1, 1990

[8] J.P.M.Hendriks: A Formalism for Speech Database Access, Speech Communication, vol. 9, No. 4, August 1990, ppg. 381 - 388

[9] THE PRINCIPLES OF THE INTERNATIONAL PHONETIC ASSOCIATION, 1949 (Reprinted 1984), International Phonetics Association, London, 1949

[10] M. Karjalainen, T. Altosaar: An Object Oriented Database for Speech Processing. Eurospeech 93, Berlin

[11] L. Kuffer: Der Einsatz eines relationalen DBMS zur Verwaltung von segmentierten und etikettierten Sprachsignal-Daten. Magisterarbeit, Institut für Phonetik, Universität München, Nov. 1991

[12] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, K. Shikano: ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis. Speech Communication, Vol. 9, No. 4, 1990

[13] D. Stott Parker: Stream Data Analysis in Prolog. In: L. Sterling, The Practice of Prolog, MIT Press, Cambridge MA, 1990

[14] B. Pompino-Marschall: PhonDat Daten und Formate. Institut für Phonetik, Uni München, 1992

[15] SAM: Assessment, Methodology and Standardisation in Multilingual Speech Technology. Int'l Symposion on Coordination and Standardisation of Speech Database and Assessment Techniques for Speech Input/ Output, Nov. 1993

[16] Chr. Saßenrath: Entwurf und Implementierung eines datenbank-gestützten Verwaltungssystems für Sprachanalysedaten, Diplomarbeit Uni Erlangen, 1991

[17] F. Schiel: An automatic segmentation program based on HMMs (working title), internal report, to appear.

[18] D. Searls: Signal Processing with Logic Grammars. TR Paoli Research Center, Unisys Co., 1989

[19] St. Seneff, V. Zue: Transcription and Alignment of the TIMIT Database, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, 1988

[20] H. Tillmann, M. Hadersbeck, H. Piroth, B. Eisen: Development and Experimental Use of PHONWORK, a New Phonetic Workbench, ICSLP 1990

Domains and Properties of Lexical Stress in German

Andreas Mengel
Institut für Kommunikationswissenschaft, Technische Universität Berlin

Abstract
This investigation addresses acoustical aspects of lexical stress in German. Different properties, units of words and methods of comparison have been evaluated in order to identify speech signal properties relevant for stress perception. It is argued that duration is the key property of stress. An explanation of how duration integrates information structure and perceptual prominence is given.

1 Motivation

In trying to identify acoustical correlates of word stress, fundamental frequency (f_0), energy and duration are usually considered as stress properties. Duration and f_0 (Heuft & Portele 1994) or only duration (Jessen et al. 1995) have been found to be dominant cues in German. These results were obtained with a set of specially designed words or sentences.

There are a couple of motivations for this study. First of all its aim is to replicate the above findings by using a larger corpus. The second reason is to evaluate possible parameters that are needed to model the perception of lexical accent: If it is duration and fundamental frequency that predict stress location best, then how do speakers of German perceive stress? A third reason for this study is to understand the functional aspects behind the signal properties of German lexical stress.

2 Specific Questions

For the purpose of this study, a set of four questions were addressed, for not only the problem of what acoustical parameters are relevant for lexical stress in German may be important but also how they are measured.

a) If stress is perceived as being located at some special position in a word, then it must be possible to identify units that allow the perception of stress. It is necessary to find out what the **units** relevant for the perception of lexical stress are. Furthermore, it is important to ask which phonological, semantic, articulatory or auditory units can serve best as these units. Candidates for possible units are morphs, syllables, vowels, and vowel-onset intervals.

b) If stress can be detected in speech signals, then there must be signal properties that mark stress and that differ systematically between stressed and unstressed units. It is possible that the signal properties used for detection of stress vary but it is hypothesized that there is one predominant signal aspect that is perceived as indicating stress. Possible candidates of **signal properties** of stress are duration, f_0, and energy.

c) If special signal properties indicate stress, then these properties must be evaluated. Different ways of evaluation are possible. **Signal property measures** considered are the mean, the minimum, the maximum, the lift, the sum, the slope, and the standard deviation of the signal properties.

d) If stress can be explained as a perception process that involves the evaluation of signal properties, it is also necessary to map or to compare these signal properties to some

reference measure. First of all, there could be an internal reference, a standard which measures of units are compared to. It might also be, that units of words are compared to each other. In that case, different kinds of comparison are possible. Thus, candidates for the kind of **comparison** that is involved when detecting stress are the absolute values, ratios, differences, the position of the maximum, and a balance measure of the property measures of units of words.

3 Data

A total of 11,644 bisyllabic word utterances (10,806 stressed on the first, 838 stressed on the second syllable) representing 782 different word types were selected (16 kHz, 16 bit mono, studio recordings, read speech, 53 speakers: 27 female, 28 male). They were extracted from the data base described in Kohler (1994). No further treatment of data has been performed as every manipulation or normalization would imply the assumption of a corresponding processing model in the human listener.

4 Parameters

In order to evaluate the suitability of the parameters described above, all possible combinations of relevant parameters have to be investigated, i.e., none of the candidates of different parameter sections (units, properties, property measures, comparison) can be evaluated in isolation. Also, although not likely, each of the combinations of the parameters are possible configurations that match well with the stress location of words. Altogether, 276 parameter configurations have to be evaluated and compared to lexical stress positions.

Units

Morphological categories were lexical morphs (L {*hund, katz-, maus,...*}), free prefixes (P {*vor, nach, bei, an,...*}), bound prefixes (p {*be-, ver-, ent-, ge-,...*}), non-native derivational suffixes (S {*-ion, -al, -ät, -ur,...*}), native derivational suffixes (s {*-lich, -ung, -bar, -heit,...*}), and inflectional suffixes (f {*-e, -t, -en, -s,...*}). For this investigation, words were divided into morphs and pseudo-morphs, no morph could have more that one vowel. Morphs without vowels were assigned to the preceding morph.

Syllables were labeled according to a morphological and a distributional principle: The morphological principle is to mark syllable boundaries in front of lexical morphs (*Taustrick* [taU.StRIk] vs. *Tauschtrick* [taUS.tRIk]). The distributional principle separates intervocalic consonant clusters into subclusters of consonants occurring at the end or at the beginning of words (*Hälfte* [hElf.te], *knipste* [knIps.t@]).

Vowels (VO) and vowel onset intervals (VI) can be identified as illustrated in Figure 1.

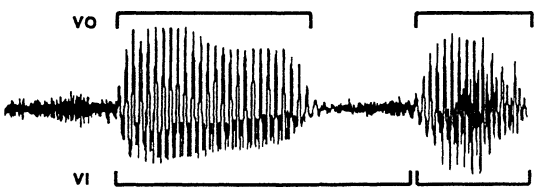


Figure 1: Vowels and vowel onset intervals.

Signal properties

Fundamental frequency was measured for voiced parts of the words. F_0 was only evaluated for sounds considered voiced. Energy was measured in RMS values of the speech signal. Both values were obtained using the function `get_f0` of the signal analysis package ESPS (Entropics). Duration was used as provided in the label data.

Signal property measures

The minimum, maximum and mean values of energy and f_0 were computed. The measure slope is the mean slope; lift is the difference of maximum and minimum of a property.

Comparisons

For this study, parameters to be compared to lexical stress positions were the absolute values of the first and the second unit, their difference, their ratio and the position of the higher value. Additionally, a balance measure was used to represent the measures of two units within a word by calculating an imaginary point of balance. Balance values range between 0 and 1 and can be used for utterances with any number of syllables.

$$M_B = \frac{1}{n-1} \sum_{i=1}^{n} \frac{v_i (i-1)}{\sum_{k=1}^{n} v_k}$$

(n: number of units; v_i, v_k: values of relevant units)

Two measurement criteria were chosen for evaluating the correspondence of parameter configurations and lexical stress positions of words. In order to account for the fact that the number of words stressed on the first syllable is ten times higher than those stressed on the second, the first criterion is a recognition design using all data: Does the absolute comparison of two parameters predict the lexical stress position? For the second measurement criterion ten selections of each 1,000 randomly selected word utterances (500 first-syllable stressed and 500 second-syllable stressed) are used and investigated computing Pearson correlations.

5 Results

The following figures give results across selected parameters that were compared while other parameters being held constant:

Figure 2 shows results for different units for duration and the balance measure. The duration of morphological units corresponds best to the variation of stress positions. In the recognition design, vowel onset intervals are second best, followed by vowels.

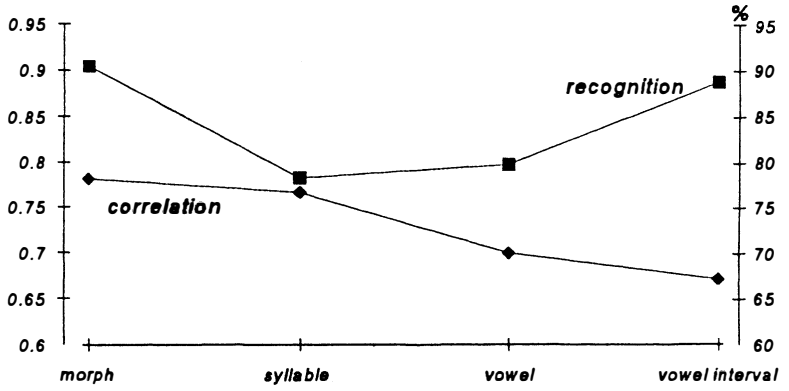


Figure 2: Correlation and recognition results of four different subunits of words using duration and the balance comparison.

Figure 3 displays results for duration, mean energy and fundamental frequency of vowels. It can be seen that duration predicts the location of stress best.

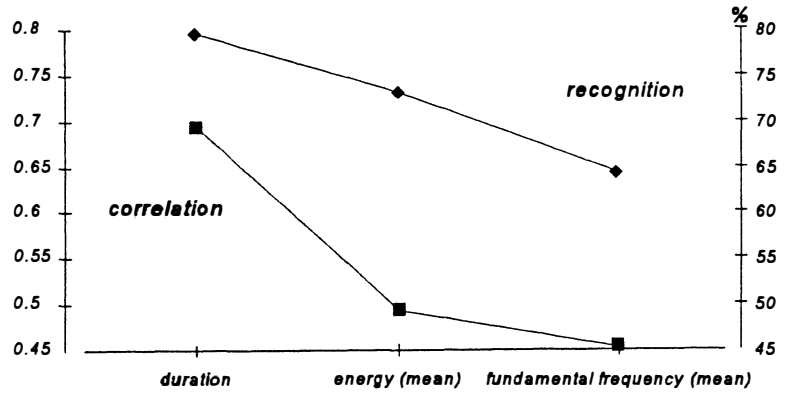


Figure 3: Correlation and recognition results of three different properties using vowels and the balance comparison.

Figure 4 shows results for f_0 using the parameters vowel and balance.

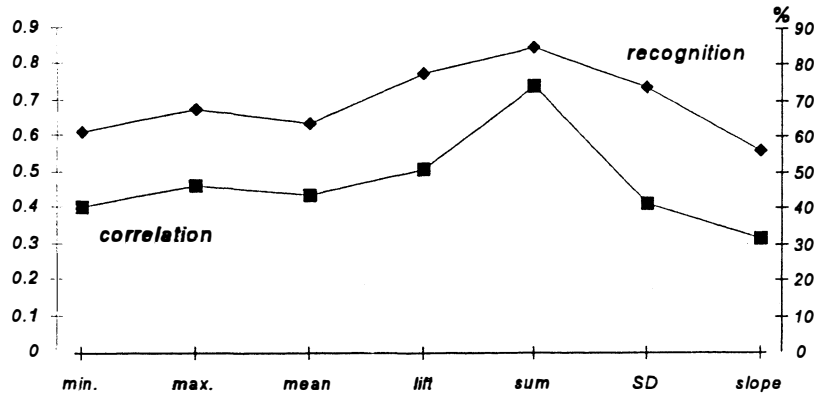


Figure 4: Correlation and recognition results comparing seven measures of fundamental frequency of words using vowels and the balance comparison.

Using fundamental frequency as a predictor of stress in German words yields recognition rates up to 85.539 % (sum of f_0). Also, the lift and SD measure are useful for stress detection.

The results obtained using different measures of energy are shown in Figure 5.

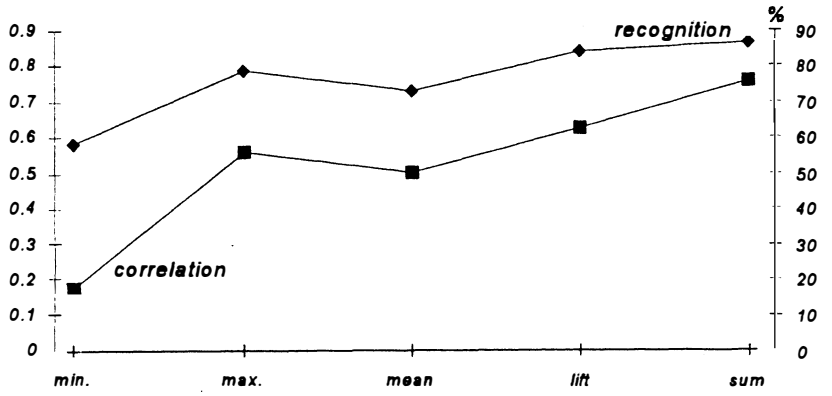


Figure 5: Correlation and recognition results comparing five measures of energy using vowels and the balance comparison.

Recognition rates of the sum of energy are even better than those for duration and fundamental frequency, also the maximum of energy can detect stress positions quite well.

Figure 6 shows the results for different methods of comparison for the property duration and the unit vowel. For this evaluation, only correlations could be computed.

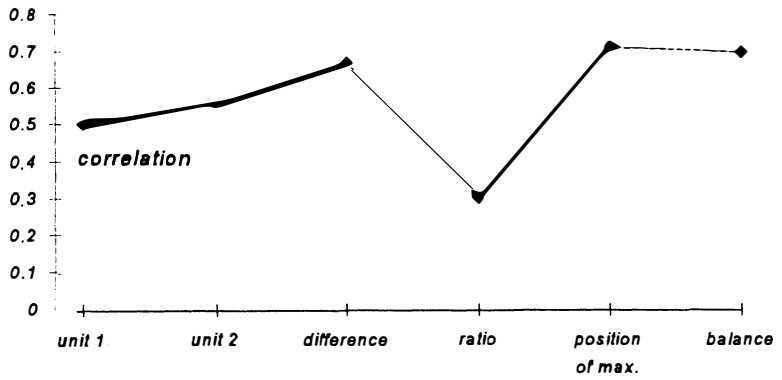


Figure 6: Correlation results for different methods of comparison units using duration and vowels.

The second unit - here the vowel - correlates more with the stress position than the first one. The best results for the measurement of stress positions can be obtained by using the positional coding, the balance value and the difference.

In order to check whether these results are valid across different linguistic and acoustical environments, some additional control measurements have been conducted: Correlations and recognition rates of five parameter combinations (positional coding of duration, mean f_0, lift of f_0, mean energy, and lift of energy) were calculated for median split subgroups of the data. The following subgroup criteria were selected: sentence position, sentence duration,

focus probability, word duration, relative word duration, speaking rate, distance to last lexical stress, sex of speaker, part of speech. For none of these subgroups there was a significant change in the predictive power of duration.

6 Conclusion on Parameter Evaluation

This investigation has shown that duration-related measures (duration, sum of energy, sum of f_0) are the most reliable measures for stress location. The most useful methods of comparison are the positional and the balance measure. The most reliable units are morphs and vowel onset intervals. Altogether, the best parameter combination for the prediction of stress positions of German bisyllabic words is the balance of the sum of energy of morphs (correlation: 0.809, recognition: 93.233 %).

The fact that RMS predicts stress positions even better than duration alone, might also hold on perceptual level: Duration perception can be manipulated by energy (Turk & Sawusch 1996), louder events are perceived as being longer. Further research has to show to what extent the better results of the sum of energy corresponds to duration perception by the listener when detecting stress positions.

No perception oriented explanation can be given to account for the finding that morphs are the best predictors of stress locations: Morphs are semantic units, there are little acoustical cues that mark morphological boundaries. Vowels and vowel onset intervals are more plausible perceptual units, both of them being well detectable because of a rise of energy at the onset (Pompino-Marschall 1989, Janker 1997).

The model of lexical stress suggested here is as follows: A rise of energy in the (speech-) signal is perceived as marking a new event. In speech signals this rise of energy coincides with the onset of vowels. Durations of these events are compared within groups of events, the longest of which is perceived as most prominent and thus stressed (cf. Figure 7).

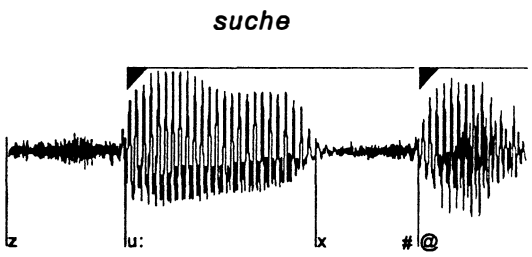


Figure 7: Vowels and vowel onset intervals.

On the signal level, morphological units show the best results for the correlation and the detection of stress location. In German, lexical stress and semantic structure correspond. On the perceptional level, vowels or vowels onset intervals (rise of energy at onset) are the more likely stress units.

What is the link between morphological and perceptually relevant units in German words? How can acoustical and semantical properties of German words be integrated in one model of lexical stress? Why is the morphological structure of German words a good predictor of lexical stress although morphological boundaries are not relevant for stress perception?

7 Morphology

Morph categories have different segmental structures: For each morph category, mean durations of morph onset, nucleus, and coda were calculated (Figure 8).

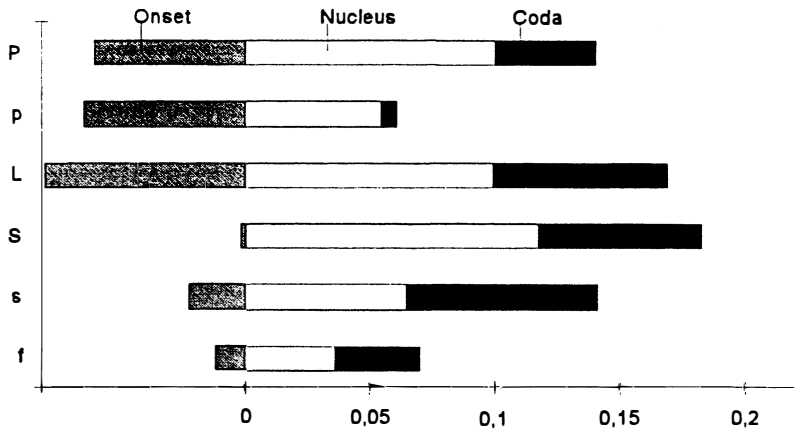


Figure 8: Onset, nucleus, and coda durations of morph types

The information of Figure 8 combined with the knowledge on morphological structures of German words can explain the correspondence of informational relevance and perceptual prominence of morphological units in German: Computing vowel and vowel onset interval durations of common word structures (L+f - *Kinder* [kInd6], L+L - *Haustür* [haUsty:6], L+s - *glücklich* [glYklIC], p+L - *bereit* [b@RaIt], P+L - *Vorzug* [fo:6tsu:k], L+S - *polar* [pola:6], the stress locations can be predicted correctly.

8 Conclusion

From the above results it can be concluded that the perception of German lexical stress is associated to durational differences of units of words. This change of duration of different units of words is due to segmental complexity: Stressed units have longer sounds and more sounds. Having more sounds has functional reasons: Morphs bearing most information (lexical morphs) are open class morphs and thus need more complex signal and segmental representations, too. Thus, German lexical stress is a function of semantic weight. During an utterance, no additional effort by the speaker is needed. Inherent signal properties of segmental compositions of morphs and the morphological structure of words evoke the stress location perceptions coded in transcription dictionaries.

References

Heuft, B.; Portele, T. (1994): Zur akustischen Realisierung des Wortakzents. Elektronische Sprachsignalverarbeitung 95: 197-204.

Jessen, M.; Marasek, K.; Schneider, K.; Clahßen, K. (1995): Acoustic Correlates of Word Stress. ICPhS 95, 4: 428-431.

Kohler, K.J. (ed.)(1994): Phonetisch-Akustische Datenbasis des Hochdeutschen. Kieler Arbeiten zu den PHONDAT-Projekten 1989-1992. AIPUK 26.

Janker, P.M. (1997): Evidence for the P-Center Syllable-Nucleus-Onset Correspondence Hypothesis. ZAS Papers in Linguistics 7: 94-124.

Pompino-Marschall, B. (1989): On the Psychoacoustic Nature of the P-Center Phenomenon. Journal of Phonetics 17: 175-192.

Turk, A.E.; Sawusch, J.R. (1996): The Processing of Duration and Intensity Cues to Prominence. JASA 99 (6): 3782-3790.

The nuclear accentual fall in the intonation of Standard German

Ralf Benzmüller and Martine Grice
Institute of Phonetics, FR 8.7, University of the Saarland

1. Introduction

In this paper we investigate the F0 contours of nuclear falls in standard German and explore the consequences for autosegmental-metrical accounts of German intonation. Where the nuclear syllable is a considerable distance from the end of the phrase, one can observe two major stages in the F0 descent: a sharp fall followed by a plateau which extends up to the phrase boundary. The beginning of the plateau, which we refer to as the elbow, has been found in informal observations to be at variable distances from the peak.

Féry (1993) and Uhmann (1991), in their analyses of German intonation, assume that the nuclear pitch accent is bitonal, involving a High and a Low tone (H*+L), followed by a L% boundary tone at the end of the intonation phrase. This analysis correctly predicts that the contour consists of a fall and a plateau, the plateau being a consequence of interpolation between the trailing accentual L tone and the L% boundary tone. However, following inter alia Arvaniti et al. (forthcoming), who argue that the F0 correlate of an unstarred tone in a pitch accent is expected to occur at a fixed distance from the peak corresponding to the starred tone, these accounts would not be able to capture the degree of variability informally observed in the alignment of the elbow.

Grice et al. (1996) on the other hand analyse German nuclear falls as comprising a monotonal H* pitch accent followed by a L tone which is independent of the pitch accent. This is an edge tone for the intermediate phrase (L-). They acknowledge that there is not a simple linear interpolation between the H* peak and the end of the intermediate phrase, as would be predicted by their analysis if L- were merely aligned with the phrase edge. They assume, as argued by Pierrehumbert (1980), Beckman and Pierrehumbert (1986) for English, that the L- already controls the F0 at some distance from the end of the phrase. However, they do not give any further indication of the exact domain controlled by the L- tone.

In the present study we investigate in an experimental setting whether the elbow in the F0 contour is aligned at a fixed distance from the H peak, either in absolute terms, measured in milliseconds, or in relative terms as the number of syllables after the nuclear syllable. We also investigate whether it occurs at the edge of a constituent such as the nuclear foot or nuclear word, the latter as argued for English by Pierrehumbert and Beckman (1988), based on evidence in Steele and Liberman (1987). Finally, we investigate whether the elbow is aligned with a lexically stressed syllable, in which case it would constitute a 'postnuclear' or 'postfocal' accent, in the sense of Ladd (1996) and Grice (1995) respectively.

Furthermore, since we are also not certain that the elbow is the point in the fall taken to be the correlate of a L tone, we examine the contours for other turning points which might have been taken to represent the L in the accounts where L is a trailing tone of the pitch accent.

Since we cannot be sure that the differences in the accounts of falls are not due to actual differences in the contours analysed rather than to the theoretical analyses employed, our study does not restrict itself to one utterance type only. We introduce differences in

information structure (broad vs. contrastive focus, see Uhmann 1991) and illocutionary force (assertive vs. directive), both of which potentially affect the type of falling contour used.

2. Experiment 1

A preliminary inspection of a corpus consisting of 40 paragraphs read by two speakers showed considerable variation in the elbow of the falling nuclear contours produced. However, since interpreting the f0 trace was often made difficult by the presence of obstruents in the postnuclear stretch, and since some postnuclear stretches were merely one or two syllables long, it was difficult to make generalisations about the exact location of the elbow in relation to any one structural landmark in the texts. In the experiment described in this section a controlled corpus was designed in which the postnuclear stretch was long and contained as many sonorant segments as possible.

2.1 Method & Stimuli

In the test corpus, the following factors were controlled for:

- number of syllables following the nuclear syllable (2, 3, 4, 5)
- interstress interval, i.e. number of postnuclear syllables before the next lexical stress (0, 1, 2, 3)
- number of postnuclear syllables before a word boundary (0, 1, 2)

Table 1 shows the set of 8 sentences upon which the experiment was based. The predicted position of the nucleus is the syllable orthographically represented as WOH, WOHN or MOHN.

1. Die Schüler	sollen	die	WOH	nung	en	#	be	MA	len
2.			WOH	nung	en	#	-	MA	len
3.			WOH	nung	-	#	be	MA	len
4.			WOH	nung	-	#	-	MA	len
5.		den	WOHN	WA	gen	#	-	MA	len
6.		das	WOHN	mo	BIL	#	-	MA	len
7.		den	MOHN	-	-	#	-	MAH	len
8.	haben	den	MOHN	-	-	#	ge	MAH	len

Table 1: Basic sentence types. CAPITAL letters mark lexical stress; # indicates a word boundary. For translation see footnote [1]

[1] 1. The pupils ought to decorate the flats
2. The pupils ought to paint the flats
3. The pupils ought to decorate the flat
4. The pupils ought to paint the flat
5. The pupils ought to paint the caravan
6. The pupils ought to paint the camper
7. The pupils ought to mill the poppyseed

The eight basic sentences were used to elicit read versions of the following types of utterance:

(a) statements with broad focus
(b) statements with contrastive focus
(c) commands

The text for the broad and contrastive focus stimuli was identical to the basic sentence set. Slight modifications to the beginning of the sentences were necessary in the commands. They started with "Du wirst sofort ..." (You will ... immediately).

The resulting three sets of sentences were randomized and read twice by two female speakers of Standard German (GS and JM). This resulted in 48 tokens for each subject[2].

After recording, the sentences were digitised and analysed with ESPS xwaves. The f0 traces were manually parameterised, marking not only the peak and the elbow, but also a third point between the two, referred to as L1 below, which was observable in the majority of the contours.

H# the f0 maximum at the nuclear accent
L1 the point where the slope of the fall turns from convex to concave, usually well above the baseline
L2 the elbow, or point where the baseline level is reached, defined as first major deviation from a "regression line" working backwards from the lowest f0 at or near the end of the contour, ignoring microsegmental perturbations.

The above labels were aligned with the f0 contour without examination of the speech pressure waveform. Nuclear and postnuclear syllable boundaries were aligned to the speech waveform with the help of spectrograms. After labelling, values were extracted for the fundamental frequency at H#, L1, and L2 and the duration was calculated between the start of the nuclear syllable and the f0 peak (H#), between H# and L1, and between H# and L2. Additionally the alignment of H#, L1, and L2 with a particular nuclear or postnuclear syllable (S1 - S6) was calculated.

2.2 Results

The results for f0, timing and syllable counts are presented separately.

2.2.1 F0 Analysis

A MANOVA was performed with the utterance type and sentence type as independent variables and the frequency at H#, L1, and L2 as dependent variables. The values for each speaker were analysed separately.

The f0 values for H# differed significantly across the three utterance types ($F = 102.5$; $df = 2$; $p < 0.001$ for GS; $F = 127.1$; $df=2$; $p < 0.001$ for JM). For both speakers commands have the highest H peaks, followed by contrastive focus, followed by broad focus statements (see

---

8. The pupils have milled the poppyseed

2 Two stimuli were deleted by accident after recording. So the total number of stimuli is only 94.

figure 1 and table 2). Post hoc tests (Scheffé) showed that each utterance type differs significantly from all others ($p < 0.001$ for JM; $p = 0.002$ for GS). Thus commands, contrastive focus statements, and broad focus statements constitute distinct groups as far as f0 height is concerned.

There is less variation in the f0 values at L1 than in the H# values. However, despite the spread being small, significant differences were found ($F = 4.2$; df=2; $p < 0.03$ for GS; $F = 3.9$; df=2; $p < 0.04$ for JM). A Scheffé post hoc test showed almost significant differences between the commands and contrastive sentences for speaker GS ($p = 0.058$) and a significant difference between commands and broad focus statements ($p = 0.035$) for speaker JM. The differences were not found to be between the same utterance types across the two speakers.

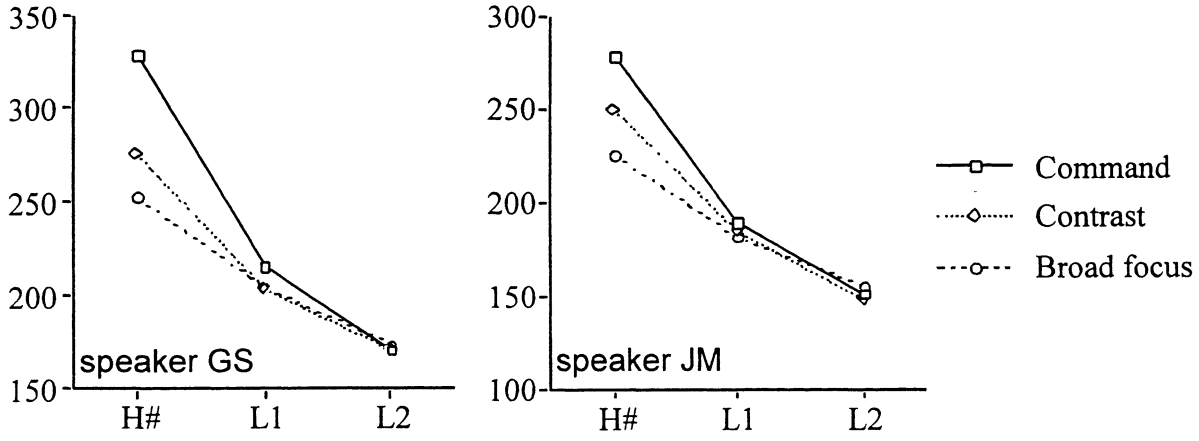


figure 1: f0 means of landmarks H#, L1, L2 in different utterance types for speakers GS and JM

	H# GS	L1 GS	L2 GS	H# JM	L1 JM	L2 JM
command	329.0 Hz (42.5)	215.7 Hz (17.8)	170.7 Hz (6.7)	278.9 Hz (15.4)	189.9 Hz (15.3)	151.5 Hz (6.1)
contrast	275.5 Hz (14.8)	203.7 Hz (9.0)	170.6 Hz (5.0)	249.5 Hz (11.7)	184.9 Hz (9.9)	148.5 Hz (3.0)
broad focus	252.1 Hz (14.8)	204.2 Hz (7.5)	173.7 Hz (3.6)	225.7 Hz (9.1)	181.2.Hz (9.7)	155.3 Hz (4.2)

table 2: means and standard deviation (in brackets) of f0 values in different types of utterance at position H#, L1, and L2 for speakers GS and JM.

The variation was even smaller for L2 values. A significant difference was found ($F = 9.7$; df=2; $p = 0.001$) only for speaker JM, the difference being between contrastive and broad focus sentences ($p = 0.001$).

Only H# values give a clear indication that the three utterance types differ consistently across the two speakers, where commands, contrastive focus statements, and broad focus statements constitute three distinct groups. For L1 and L2 there are differences which are less significant and less consistent across speakers. We now turn to timing and investigate whether there are any consistent differences in that area.

2.2.2 Timing

As for the frequency analyses, a MANOVA with certain durational measures as dependent variables, and utterance and sentence type as independent variables was carried out.

Across the 8 basic sentence types, the timing of L1 is relatively constant in relation to the f0 peak. The duration from S0 (beginning of the nuclear syllable) to H# is correlated with the duration from S0 to L1 ($r = 0.51$; $df = 2$; $p < 0.001$). No effect of sentence type and utterance type was found in duration from H# to L1 in either speaker. This points to L1 being located a fixed distance after the nuclear peak regardless of sentence or utterance type.

Duration of	Speaker	mean	st. dev.
H# to L1 (ms)	GS	121.0	32.4
H# to L1 (ms)	JM	176.9	40.6
H# to L2 (ms)	GS	352.7	108.3
H# to L2 (ms)	JM	438.4	107.8

table 3: Distance of H# to L1 and H# to L2 in ms, mean and standard deviation for speakers GS and JM.

Unlike L1, the position of L2 is highly variable as the means and standard deviations in table 3 show. Moreover it appears to be independent from H#, since the duration between S0 and H# is not significantly correlated with the duration from S0 to L2.

For H# to L2 only speaker JM showed significant differences for utterance type ($F = 3.9$; $df=2$; $p = 0.033$). The high variation in duration from H# to L2 is mainly caused by sentence type with highly significant differences ($F = 15.6$; $df=2$; $p < 0.001$ for JM; $F = 12.4$; $df=2$; $p < 0.001$ for GS). The extremes in the distance between H# and L2 were found for both speakers in the sentences with "Mohn malen" (JM mean = 304.5 ms; GS mean = 190.5 ms; interstress interval = 0) on the one hand and "Wohnungen bemalen" (JM mean = 530.7 ms; GS mean = 452.6 ms; interstress interval = 3) on the other. This indicates that L2 tends to occur later as the interstress interval is increased. However the absolute timing does not tell us which syllable L2 is aligned to. This is investigated in the next section.

2.2.3 Syllable alignment

Since absolute durational measures are speaker and speech rate dependent, and they do not give information about relations to the linguistic structure of the utterance, the location of the landmarks in the f0 contour in relation to syllables was examined.

The position of L1 is in 84.0 % of the cases one syllable after the syllable aligned with the nuclear peak. In 10.6 % of cases L1 occured two syllables after the H peak rather than one.

This was the case if the immediately postnuclear syllable was short[3], or it was part of a longer foot of 3 or more syllables[4]. In only 5.3 % it is in the same syllable.

The highly significant correlation between the distance of the peak and that of L1 from the beginning of the nuclear syllable referred to in section 2.2.2 would in fact predict that L1 is not bound to a particular syllable but rather to a set distance which happens to correspond mainly to the postnuclear syllable. A richer variety of syllable structures and weights would have to be introduced to test this hypothesis exhaustively. However the present data have not been designed to examine this question.

The syllables aligned with L2 are shown in table 4. L2 is most often (94.7 %) aligned with the syllable MA of "(be)malen". In 92 out of 94 (97.9 %) cases, L2 was aligned with a lexically stressed syllable after the nucleus[5].

Syllable	number	percent
MA (malen)	89/ 94	94.7 %
BIL (wohnmobil)	2 / 12	16.7 %
WA (wohnwagen)	1 / 12	8.3 %
len (malen)	2 / 94	2.1 %

table 4: Alignment of L2 with syllables - number and percentage

In the sentences containing "Wohnwagen" and "Wohnmobil" there is variation as to the lexically stressed syllable upon which the baseline is reached. In 16.7 % of the "Wohnmobil" sentences L2 is on BIL and in 8.3 % of the "Wohnwagen" cases it is on WA. This implies that there is a degree of choice on which stressed syllable to reach the baseline[6]. In most cases it was not the secondary stress in the compound, but the stress on the following verb which aligned with L2.

Now, if L2 is predominantly aligned with a stressed syllable, the question arises as to whether it could be interpreted as occuring at the boundary of a foot, since the stressed syllable always marks the beginning of a new foot. If the L tone is interpreted as an edge tone for the nuclear foot, then we would expect that L2 would have aligned more often with BIL of "Wohnmobil" and WA of "Wohnwagen" because these constitute the beginning of the

---

3 L1 occurs 2 syllables after the peak 3 times (out of 11) in "mohn gemahlen", twice (out of 12) in "Wohnmobil malen".

4 L1 occurs 2 syllables after the peak twice (out of 12) in "Wohnungen malen", and 3 times (out of 12) in "Wohnungen bemalen".

5 In fact, L2 is located in the stressed syllable more often than the starred tone of the pitch accent, here labelled as H# (77 out of 94 = 81.9 %)

6 In an additional recording session the two speakers read the basic sentence set twice as stylised contours. While speaker GS realised the step down always on the syllable MA of "(be)malen" (paint/ decorate), speaker JM chose WAG of "Wohnwagen" (caravan) and BIL of "Wohnmobil" (camper) as the stepdown location. Although this is a different contour, sometimes referred to as a 'stylised fall' (e.g. Ladd (1996)) the fact that the step down to mid occurred on the secondary stress of the compound words some of the time as well as on the lexical stress of the main verb supports the possibility of an alignment of postnuclear tones with lexical stress.

immediately postnuclear foot. This was in fact not the case (10/12 in Wohnmobil and 11/12 in Wohnwagen).

Additional evidence is given by the following findings: For each speaker we calculated the means of the location of L2 in the syllable MA- of "(be)malen", if it was located on that syllable. For speaker JM, L2 is located 56.2 % (154.7 ms) into the MA syllable and for speaker GS it is located 40.4 % (92.2 ms). Thus L2 is more oriented towards the center of the syllable than to its edges. There are no cases of L2 occuring immediately before the syllable MA. These findings lead us to reject the hypothesis that L2 occurs at the nuclear foot boundary or indeed at any foot boundary.

The cases where the foot and the word boundary are in different positions are particularily interesting for the question as to whether the fall is completed at the end of the nuclear word, which would have supported the analysis of the L tone as an edge tone of the word, as claimed for English. From the data shown in table 5 it is clear that word boundary does not play a role in the location of L2, because it does not occur in the vicinity of the word boundary at all.

	L2 2 or more syllables after word boundary	L2 one syllable either side of the word boundary
MOHN ge/be	11	0
WOHnung be	12	0
WOHnungen be	12	0
sum	35	0

table 5: Alignment of L2 in words where foot and word boundaries are distinct

We can conclude that L2 aligned neither with the right edge of the nuclear word nor with the right edge of the nuclear foot.

3. Experiment 2

In Experiment 1, L2 was, in the vast majority of the cases, located in the syllable MA of "(be)malen". This could be due to "(be)malen" being the last word in the utterance or to the fact that MA was the penultimate syllable. A follow-up experiment with one speaker and a reduced number of utterance types was carried out in order to test whether the final position of "(be)MAlen" led to the frequency of its alignment with L2.

3.1 Method and Stimuli

The 8 basic sentences were read by speaker GS in contrastive and broad focus statements. After "(be)malen" the word "wollen" was added and the beginning of the sentence was adjusted appropriately. The sentences had the following pattern:

"Die Schüler hatten die ... (be)malen wollen." (the pupils wanted to paint/ decorate the ...).
The empty slots were filled according to the basic sentence set in table 1. The recorded stimuli were analysed as in experiment 1.

3.2 Results

The results for the additional sentences show more variability as to the position of L2 than in experiment 1 (see table 6). L2 is in 68.8 % cases aligned with the syllable MA. In 75 % it is on a lexically stressed syllable. Despite the higher proportion of unstressed syllables aligned with L2 it is clear that neither the last word nor the penultimate syllable is the preferred location of L2. In fact it never occurred on the final word "wollen".

	broad focus		contrast	
	ma	else	ma	else
mohn	2		1	1 len
wohnwagen	1	1 WA	1	1 gen
mohn ge/be	2		2	
wohnung	1	1 len	1	1 ung
wohnmobil	2		1	1 BIL
wohnung be	1	1 be	2	
wohnungen	1	1 en	1	1 en
wohnungen be	2		1	1 en
Sum	12	4	10	6

table 6: Number of occurences of L2 aligned with syllables in two types of utterance: broad focus and contrastive sentences.

4. Discussion

We have been able to observe two landmarks after the peak, rather than one. The first, L1, is a flex point where the fall turns from convex to concave. The second, L2, is the beginning of the plateau at which the fall reaches the baseline. In section 4.1 we discuss whether differences in utterance type reflected in the height of the peak affect the timing of the fall. In section 4.2 and 4.3 we address the status of L1 and L2 as corresponding to phonological entities. Finally in section 4.4 we explore the implications of our findings for autosegmental-metrical theory.

4.1 F0 height and timing

The experiments have shown that there are consistent differences in the height of the H peak across utterance type, indicating three distinct F0 ranges for broad focus statements, contrastive statements, and commands. However, this height distinction does not appear to affect the timing of L1 and L2 in relation to the peak. Leaving open the questions as to whether the F0 height differences are local or global, and whether they are phonological or phonetic, we can assume that the phonological analysis of all three falling contours consists of the same number of tones. We have not found any differences in the contours analysed which could have resulted in the differences between Féry and Uhmann on the one hand, who incorporated a L tone into the accent (H*+L), and Grice et al. on the other who treated the L tone as independent of the accent.

4.2 The status of L1

It was found that the timing of L1 was correlated with that of the foregoing H peak. It might be argued that the L1 therefore corresponds to the L trailing tone of Féry's and Uhmann's H*+L pitch accent. However the truly phonological status of the L1 correlate is unclear. There might be an entirely phonetic explanation for the presence of a discontinuity in the F0 descent. We might for instance take it to be a reflection of the fact that relaxation and contraction of different sets of muscles are involved at different stages in the fall. Assuming that the first part of the fall is achieved by relaxing the crycothyroid muscle, which is known to have a raising effect when contracted, the flex point could mark the stage at which an equilibrium is reached, after which further lowering is achieved only by the contraction of lowering muscles such as the sternohyoid (cf. Strik & Boves 1992, Beckman et al. (1995), and Erickson et al. (1994)). If this is the case, then we would not necessarily expect L1 to occur at a fixed time from the peak, regardless of the distance in Hz between the two points, unless the increased tension needed to produce higher peaks leads to a faster return to the equilibrium level. This issue requires further investigation.

4.3 The status of L2

The elbow of the fall, referred to as L2, was shown, as predicted from our preliminary investigations, to occur at variable distances from the peak, depending on the text of the sentences. In their discussion of English intonation, Pierrehumbert and Beckman (1988) suggest that the intermediate phrase edge tone is simultaneously an edge tone for the nuclear word. In the present study of German, we found no evidence in favour of treating L2 as the correlate of a word-final edge tone. In fact, in cases where word boundaries did not coincide with other boundaries, L2 was, with only two exceptions which occurred in experiment 2, at least two syllables away from the word boundary. The only plausible candidate for a boundary aligning with L2 was the right edge of a foot. However, there were two indications which led us away from treating L2 as the correlate of a foot-final edge tone. First, in the two sentences containing nominal compounds 'Wohnwagen' and 'Wohnmobil', L2 occurred in the majority of cases near the edge not of the nuclear foot, but of a postnuclear foot. If we were

positing a foot-final edge tone, the foot concerned would not be the nuclear one, neither would it be the final in the phrase, as experiment 2 has shown. Second, L2 usually occurred not at the boundary but well into the first syllable of the subsequent foot. It also rarely occurred before the boundary. This indicates that it is the stressed syllable itself which attracts L2 alignment.

Whereas in the sentences with nominal compounds there appears to be a choice as to where to align the L2 - supported by data on stylised stepdown contours[7] - it never aligned with the stressed syllable of the auxiliary verb 'wollen' in experiment 2. Thus semantic weight appears to be involved in the choice. All this - alignment with stressed syllable, speaker choice, and semantic weight - points to an accentual function of L2. However, the accentuation brought about by the L tone is clearly secondary, as it does not affect the primary focal structure, which means that the sentence stress remains unchanged. This secondary accentuation is akin to what Kohler (1996) includes in his German text-to-speech synthesis system as 'partial deaccenting', where postnuclear as well as prenuclear lexical stresses are not fully deaccented but rather receive a small degree of accentuation.

4.4 Implications for autosegmental-metrical theory

The accentual function of the L tone poses problems for the established positional definition of the nucleus as the last accent in the phrase (cf. Pierrehumbert 1980). However, German and English are not the only languages requiring a redefinition of the nuclear pitch accent. As Ladd (1996) points out, similar conclusions have been drawn for other languages which have been shown to have accents following the sentence stress: eg. Palermo Italian (Grice 1995), Maltese (Vella 1994), and Portuguese (Frota 1997). Solutions to this problem have been proposed by Grice (1995) and Ladd (1996). Grice associates accents to phrasal nodes in a prosodic tree in which the designated terminal element receives the nuclear accent, after it has percolated down along the strong branches. The nuclear accent is thus not the final accent in the phrase, but the accent which is associated with strong nodes only. Ladd accords the nuclear accent a special status and permits a following tone which may have accentual properties (given an accentable item in the postnuclear text), or may simply surface at the phrase boundary. He refers to this tone as a 'phrase accent'.

In the data reported on here, the L tone corresponding to L2 clearly resembles the phrase accent, in that it has a dual function: it is simultaneously a boundary tone for the intermediate phrase and a secondary accent. Although the issue of whether the accent should be represented as H* or H*+L is left open, the results presented here support the analysis of the fall as having a low target independent of the nuclear accent and suggest a complex function of this tone.

---

[7] cf. footnote 5

5. References

Arvaniti, Amalia, D. Robert Ladd and Ineke Mennen (Forthcoming): Stability of Tonal Alignment: the Case of Greek Prenuclear Accents, Journal of Phonetics.

Beckman, Mary E. and Janet B. Pierrehumbert (1986): Intonational Structure in English and Japanese. Phonology Yearbook 3: pp. 255-310.

Beckman, Mary, Donna Erickson, Kiyoshi Honda, Hiroyuki Hirai, and Seiji Niimi (1995): Physiological correlates of global and local pitch range variation in the production of high tones in English. in: Proc. ICPhS, Stockholm.

Erickson, Donna, Kiyoshi Honda, Hiroyuki Hirai, Mary E. Beckman, and Seiji Niimi (1994): Global pitch range and the production of low tones in English intonation. in: Proc. III. ICSLP, Yokohama, pp 651-654.

Féry, Caroline (1993): German Intonational Patterns. Tübingen.

Frota, Sonia (1997): Association, Alignment and Meaning: the Tonal Sequence HL and Focus in European Portuguese. In: Proceedings of the ESCA Workshop on Intonation, Athens, pp. 127-130.

Grice, Martine (1995): The Intonation of Interrogation in Palermo Italian: Implications for Intonation Theory. Tübingen.

Grice, Martine, Matthias Reyelt, Ralf Benzmüller, Jörg Mayer and Anton Batliner (1996): Consistency in Transcription and Labelling of German Intonation with GToBI, Proc. IV. ICSLP, Philadelphia, pp. 1716-1719.

Kohler, Klaus (1996): Parametric Control of Prosodic Variables by Symbolic Input in TTS synthesis. In: van Santen et al. (eds.) Progress in Speech Synthesis, Springer, New York, pp. 459-475.

Ladd, D. Robert (1996): Intonational Phonology. Cambridge University Press, Cambridge.

Pierrehumbert, Janet B. (1980): The Phonology and Phonetics of English Intonation. MIT Cambridge, Bloomington.

Pierrehumbert, Janet B. and Mary E. Beckman (1988): Japanese Tone Structure. Cambridge, London.

Steele, S. and Mark Liberman, (1987): The shape and alignment of rising intonation, JASA 81, Supl. 1, p. 79.

Strik, Helmer & Lou Boves (1992): A physiological model of intonation. in: Proc. of the Dept. of Language and Speech, University of Nijmegen, Vol. 16/17, pp. 96-105.

Uhmann, Susanne (1991): Fokusphonologie. Niemeyer, Tübingen.

Vella, Alexandra (1994): Prosodic Structure: Intonation in Maltese and its Influence on Maltese English, PhD dissertation, University of Edinburgh.

Accounting for the phonetics of German *r* without processes

Adrian P. Simpson
Institut für Phonetik und digitale Sprachverarbeitung
der Christian-Albrechts-Universität zu Kiel

1 Introduction

A rich variety of phonetic patterns are associated with German **r**. Leaving aside interdialectal and interstylistic differences, these patterns are determined by the place in the syllabic and rhythmic structure. For a typical North German speaker the correlates of **r** can range from a voiceless uvular fricative in the initial consonant cluster of a word such as *trat* ("stepped") to apparent absence of anything at all following the open vowel in a word such as *Bart* ("beard"). In particular, the patterns observed for **r** in postvocalic position are particularly rich. In combination with short quantity vowels (e.g., *wird* "becomes, will", *Wurst* "sausage", *Korb* "basket", *Erna* proper name) we can find phonetically long monophthongs, whose quality is opener and more central than the quality of their **r**-less congeners. In combination with non-open long quantity vowels (e.g., *ihr* "her", *Uhr* "clock", *wer* "who") we can find phonetically long diphthongs which end centrally somewhere between [ə] and [ɐ].

Both in extensive descriptive surveys (Ulbrich 1972; Graf and Meißner 1996) as well as phonological analyses (Hall 1993), these patterns are accounted for in terms of generative processes. The names of the descriptive categories 'vokalisiert' and 'elidiert' used in Ulbrich (1972) and Graf and Meißner (1996) are process-oriented. In Hall's lexical phonological account all allophonic variants are derived from the consonantal specification of a voiced uvular trill employing rules such as '[ʀ]-vocalisation' (Hall 1993: 88).

There would appear to be both formal linguistic as well as phonetic grounds why a generative phonological account of the phonetic patterns associated with German **r** is inappropriate. In non-linear approaches to accounting for phonetic patterns such as Firthian phonology (Firth 1948) and more recently in related declarative frameworks (Coleman 1994; Local and Ogden 1997) as well as articulatory phonology (Browman and Goldstein 1989), it has been successively shown that differences in the phonetic appearance of the same phonological objects can be accounted for using a combination of rich phonological structure and non-linear phonetic exponency, avoiding the need for the destructive might of rewrite rules. From a phonetic point of view a different interpretation of the cases of elision reported in Ulbrich (1972) and Graf and Meißner (1996) suggest that the phonetic correlates of the phonological object **r** are not absent.

Using a declarative phonological analysis and non-linear phonetic exponency this paper demonstrates that the complex set of consonantal and vocalic patterns associated with **r** can be reduced to two phonetic exponency statements, without using processes such as vocalization and elision. One exponency statement describes the consonantal correlates associated with **r** at syllable onset. The second exponency statement describes the vocalic correlates of **r** at coda. The monophthongal vocalic qualities found in connection with short quantity vowels as opposed to diphthongal patterns found with long quantity vowels are seen as the product of differences in co-temporality of the correlates of **r** and those of the vowel. However, these differences are not treated as being specific to **r**, but rather as part of the more general observation that consonantal strictures following short vowels are often longer than those following long vowels. Verification is provided using synthesis examples produced by a computer implementation of the phonological abstractions and their phonetic exponents.

2 The phonetics of German *r*

Comparing descriptions from around the end of the last century (Bremer 1893; Viëtor 1894) with contemporary analyses (Kohler 1995), it is striking how little has changed in the complex set of phonetic patterns associated with **r** in Standard German.

The consonantal correlate of **r** is a dorso-uvular stricture of close or open approximation. The state of the glottis accompanying the stricture can be open, narrowed or vibrating. The state of the glottis and the type of stricture are partly contextual and partly speaker-specific. In voiceless plosive and fricative onsets the glottis is open and a dorso-uvular stricture of close approximation gives rise to friction. In other onsets and intervocalically the glottis is ready for voice[1], but if the dorso-uvular stricture is too small, the build-up of air pressure between the glottis and the supraglottal stricture can be sufficient to suppress vocal fold vibration (Bickley and Stevens 1986, 1987; Stevens 1987).

Figure 1 contains sonagrams and annotations of utterance portions illustrating these different consonantal possibilities[2]. Example 1(a) is from a male speaker, the remaining examples are from female speakers. The portion labelled with **$r** in each case is of interest. The first three examples (1a-c) show voiceless uvular friction from voiceless plosive (1a) and fricative (1b-c) onsets, taken from the words (a) *Eintracht* ("harmony"), (b) *schreiben* ("write") and (c) *Freitag* ("Friday"). The uvular friction in each case is characterized by strong excitation of F2-F4, most clearly visible in the female examples (1b-c). Figure 1(e-f) illustrates unvoiced (e) and voiced (f) uvular friction in two tokens of the proper name *Doris* uttered by the same speaker. As both tokens are temporally very similar and were produced by the same speaker, the presence or absence of vocal fold vibration would appear to arise solely from differences in the size of dorso-uvular stricture. The most complex glottal activity during the uvular stricture arises in lenis plosive onsets. The utterance portion annotated with **$-h** and **$r** in 1(d) is a typical example. Following the release of the velar plosive the fricative stricture is unvoiced, then after about 30 ms voicing begins, only to be almost completely suppressed again after a further 20 ms. Similar complexity can also be found following labial and apical plosives. This complexity arises from the instability of uvular strictures, which are particularly susceptible to abrupt changes in air pressure and flow occurring directly after plosive release.

At coda the phonetic correlate of **r** is a central half-open vowel quality. However, in the majority of cases this quality is not temporally delimitable in the acoustic record in the same way as the dorso-uvular fricatives and approximants just described. Instead we find a range of monophthongal and diphthongal vowel qualities, which are temporal amalgams of the phonetic correlates of **r** and those of the vocalic nucleus. Monophthongal [ɐ] for the weak syllable **ər** is well-known from the literature (Meinhold 1989; Kohler 1990; Kohler 1995; Barry 1995), but in combination with other vowels we are led to expect diphthongal vowel qualities which begin at the quality of the **r**-less vowel and move towards a [ɐ]-quality. The descriptive dichotomy of a monophthongal [ɐ] for **ər** and diphthongs in combination with other vowels is largely based on observation of isolated words and syllables and oversimplifies the patterns found even in

[1] Cf. Lisker and Abramson 1964, p. 415: 'If the speaker closes the glottis down enough for phonation, he does not directly "command" the vocal folds to vibrate; rather, he makes the necessary muscular adjustments that set the conditions for vibration when sufficient airflow is supplied.'

[2] All the examples in this paper are taken from speakers who produced the Marburg and Berlin sentence set from the *Kiel Corpus of Read Speech* (Kohler, Pätzold, and Simpson 1995; IPDS 1994). The index of each example (e.g. k08mr074) refers uniquely to an utterance by a particular speaker from a particular subcorpus. The index 'k08mr074' in Figure 1, for instance, refers to sentence 074 of the Marburg sentence set spoken by speaker k08 (uneven numbers are male speakers, even female). The natural and synthetic utterances contained in the figures and tables in this paper can be found at the following URL: `www.ipds.uni-kiel.de/examples.html`.

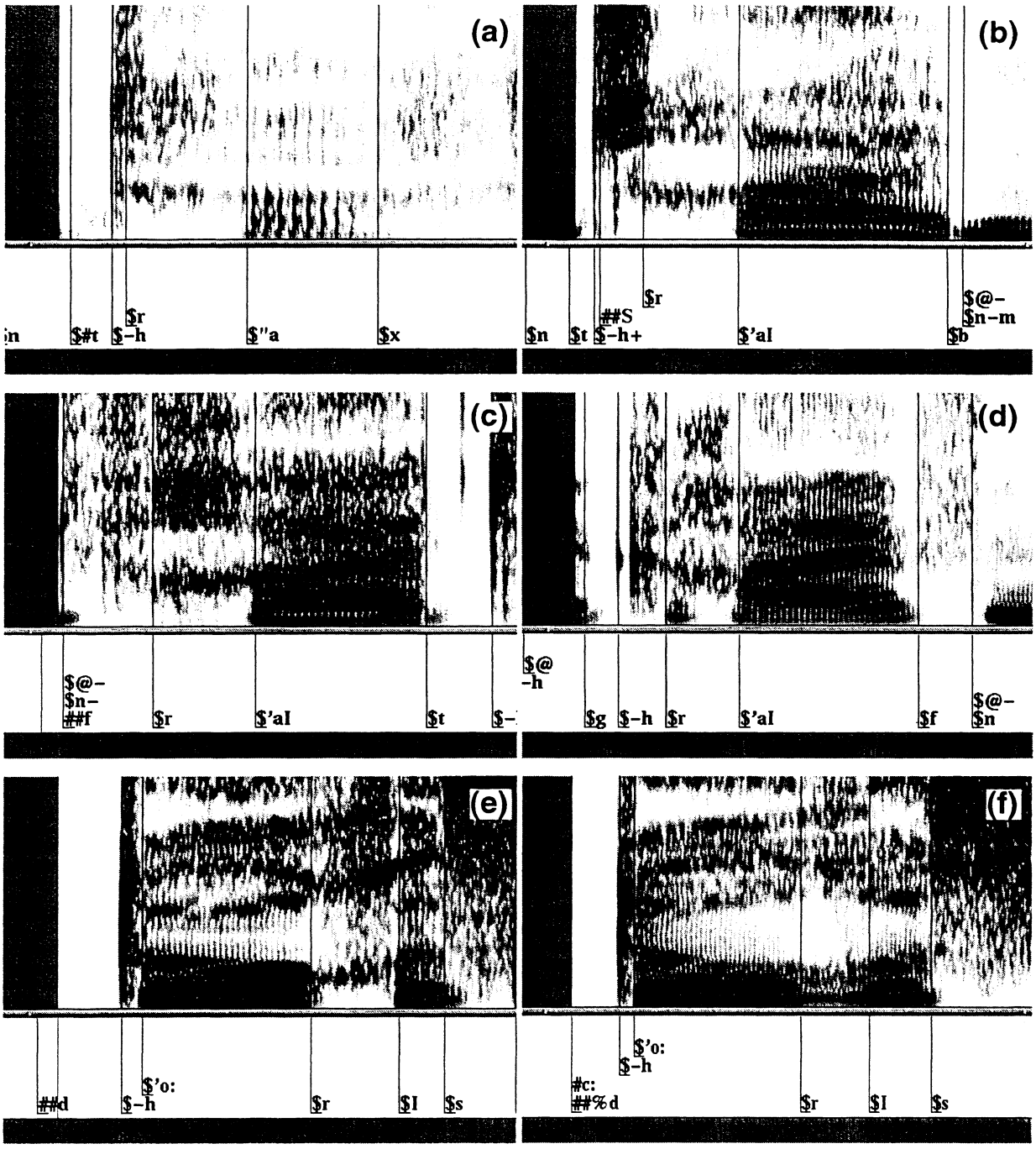


Figure 1: Sonagrams and annotations illustrating different strictures and states of the glottis associated with the dorso-uvular correlate of **r**. Examples (a-c) are voiceless uvular fricatives found in voiceless plosive and fricative onsets, (d-f) are from other onset and intervocalic positions (see text). Examples (b-f) are from female speakers, (a) from a male speaker. (Refs.: (a) k07mr055, (b) k10mr095, (c) k10mr063, (d) k12mr027, (e) k08mr026, (f) k08mr074)

laboratory read speech. In long quantity syllables the phonetics of the vowel and **r** give rise to diphthongs beginning at a quality akin to the **r**-less vowel and ending centrally between [ɔ] and [ɐ]. In short quantity syllables the phonetics of the vowel and **r** produce monophthongal or slightly diphthongal vocalic portions, whose quality is open and central of the corresponding **r**-less vowels.

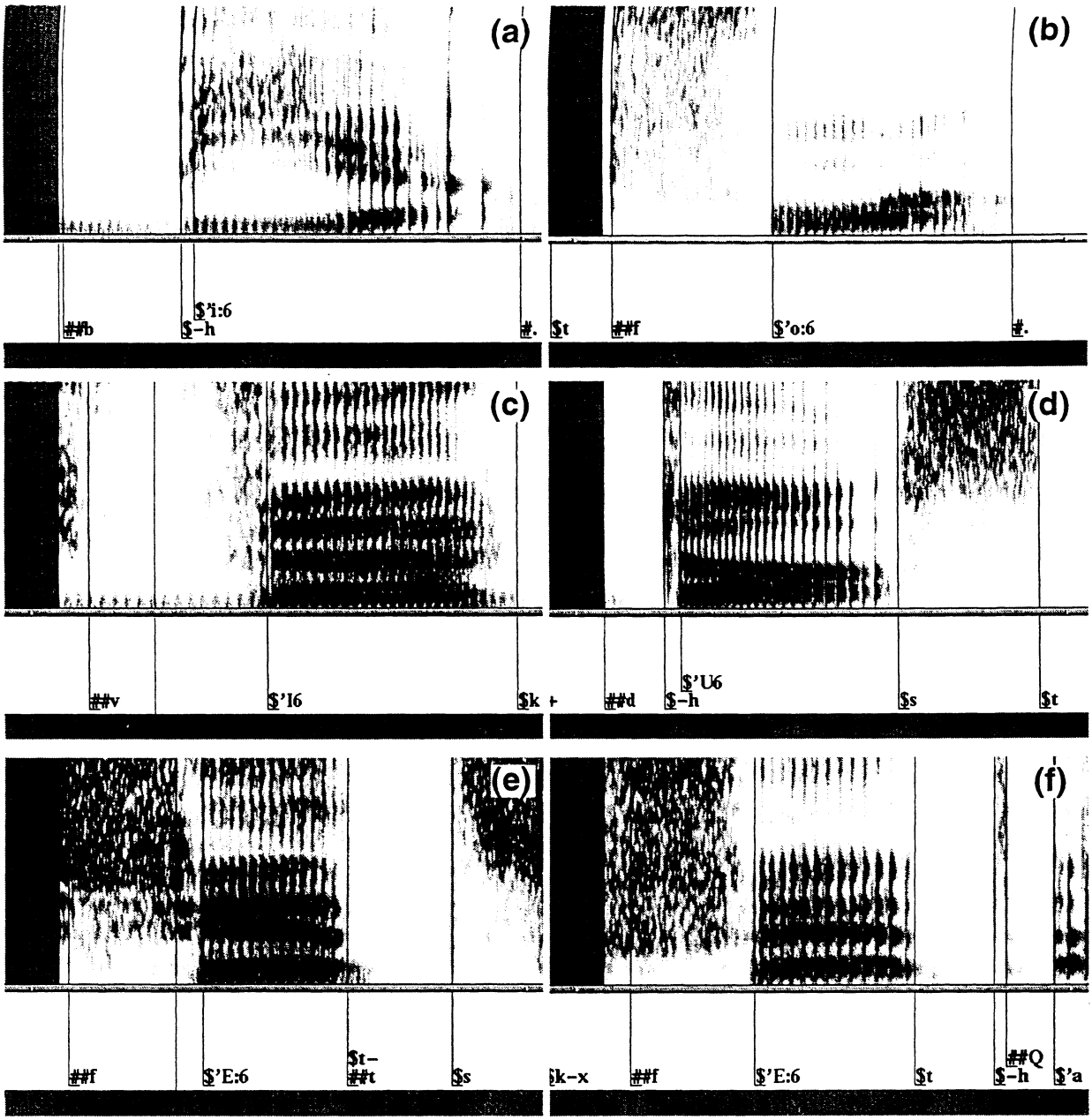


Figure 2: Sonagrams and annotations of a selection of short and long quantity **r**-vowels produced by the male speaker k67. Utterance portions are from the words (a) *Bier* ("beer"), (b) *vor* ("before"), (c) *wirklich* ("really"), (d) *Durst* ("thirst"), and (e-f) *fährt* ("goes, drives"). (Refs.: (a) k67mr089, (b) k67mr058, (c) k67mr090, (d) k67mr062, (e) k67mr026, (f) k67mr071)

Figure 2 contains sonagrams and annotations of short and long quantity **r**-vowels. Figure 2(a) and (b) are examples of long quantity vowels, 2(c) and (d) short quantity. The long quantity vowels in 2(a) and (b) are both utterance final and the diphthongal formant movements from the vowel space periphery to a central half-open position in both cases is clearly visible. Indeed, the beginning of the syllabic portion in 2(a) is voiced dorso-palatal friction. In stark contrast to the clear diphthongs in 2(a) and (b) are the short quantity vowels in 2(c) and (d). The vocalic portions here are also relatively long, ca. 150 ms for *wirklich* ("really") and 200 ms for *Durst* ("thirst"). However, in both cases the auditory quality of the vocalic portions is monophthongal, which is reflected in the presence of any significant formant movements only in the transitional

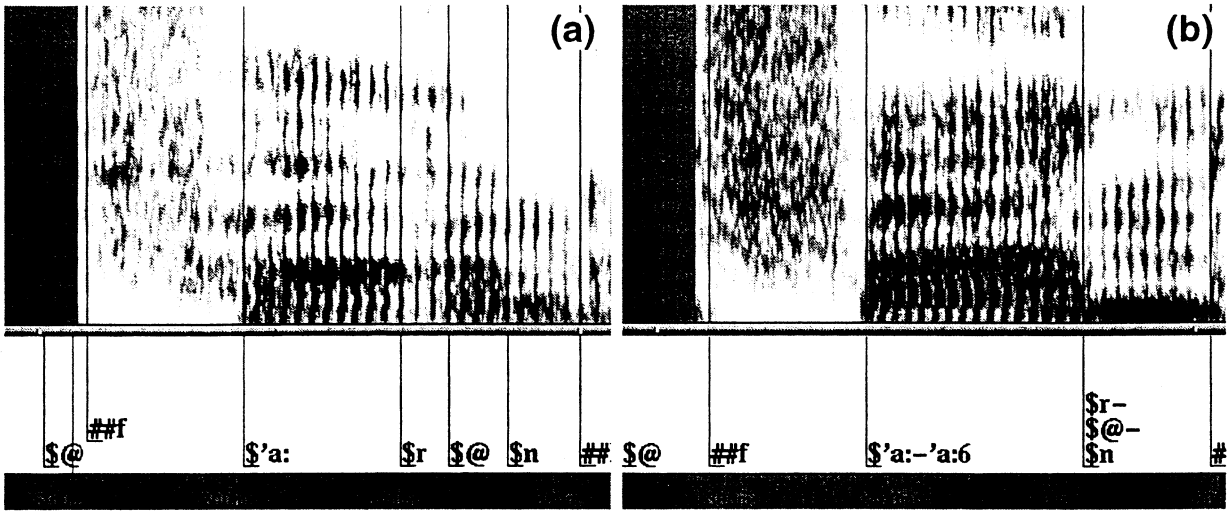


Figure 3: Sonagrams and annotations of (a) consonantal and (b) vocalic tokens of the verb *fahren* from two male speakers. (Ref.: (a) k07mr088, (b) k11mr088)

periods away from and into adjacent consonants.

Two further aspects complicate the description of **r**-vowels. First, different tokens of the same word by the same speaker can have different vowel qualities. Tokens of the word *fährt* ("goes, drives") in Figure 2(e) and (f) illustrate one such example. The word *fährt* occurs twice in the Marburg sentence set, in *Doris fährt zu weit links* ("Doris is driving too far to the left.") and *Vorsicht, Zug fährt ab!* ("Mind out, the train is departing!"). Consistently across the twelve speakers who produced this sentence set, the vocalic portion of the first token is qualitatively closer and more likely to be diphthongal than that of the second token. It is not clear on the basis of the data base material used whether this is categorial ambivalence or due to long domain vowel harmony. The systematic nature of the difference points towards harmony. In the first case the vowel of *fährt* is surrounded by vowels which are half-close and close in quality, whereas in the second sentence the *fährt* is followed by the open vowel of the verbal particle *ab*. The second complication are differences in the quantitative distribution of **r**-vowels across different lexical and grammatical items. So, for instance, the grammatical item *wer* ("who") for different speakers can have either the half-open quality of the short quantity vowel or the half-close and markedly diphthongal quality of the long quantity vowel.

In the majority of cases the distribution of consonantal and vocalic correlates of **r** is clear-cut. However, the phonetic shape of certain lexical items can alternate between between the vocalic and consonantal correlates. This alternation most commonly occurs in a **Vrən** configuration found primarily in certain verb forms (e.g. *fahren* "go, drive") and plural nouns (e.g. *Erdbeeren* "strawberries"). Figure 3 shows (a) consonantal and (b) vocalic tokens of the verb *fahren* from the same sentence uttered by two male speakers (k07 and k11). In 3(a) vocalic portions are visible on either side of the voiced dorso-uvular fricative. In 3(b) an open vocalic portion extends from the labiodental friction to the onset of the final nasal. However, both tokens have approximately the same duration, and are both disyllabic. In 3(b) both the nasal and the open vocalic portion are longer and the nasal is syllabic. In the next section structural differences will be proposed to account for the consonantal and vocalic tokens, but it is unclear whether a speaker chooses to produce one or the other variant purely on stylistic grounds or whether structural ambivalence may also play a part.

3 Accounting for the phonetics

Recent attempts at accounting for the complex patterns associated with German **r** have been in generative phonological terms. This applies not only to generative phonological analysis proper (Hall 1993), but also to Ulbrich's (1972) extensive descriptive survey. In the introduction both phonetic and phonological difficulties with a generative approach were identified.

The phonetic problem is one of data interpretation which is not exclusively generative, but is undoubtedly nurtured by generative formalism. It can best be illustrated using the descriptive categories which Ulbrich uses to classify vocalic allophones. Ulbrich analyses some 11000 /r/ allophones taken from recordings of news broadcasts, programme announcements and literary texts produced by 25 radio announcers and 15 actors. Each allophone is classified according to auditory and structural criteria. Five primary articulatory categories are proposed ("r-trills", "r-fricatives", "r-vowels", "r-elision" and "r-indifferent"). The first three of these categories are then further subdivided. So, for instance, "r-fricatives" is divided into [ʁ, ɣ], [ɹ] und [χ, x].

The most problematic aspect of Ulbrich's analysis is his categorization of vocalic allophones. Whereas the classification of consonantal allophones pays attention to articulatory and phonatory detail the categorization of the vocalic allophones is coarser and in places confusing. The "r-vowels" category is divided into [ɐ] and [V^{ɐ}] for monophthongal [ɐ] cases and diphthongs, respectively. Given the enclosure in [ ] we would expect the classification of a vocalic allophone as [V^{ɐ}] to mean a diphthong whose quality ends at a half-open central position, but the description in places shows this not to be the case, e.g.

> ... *durch* zwischen [dʊɐç] und [dʊʁç].
> ([ɐ] tendiert in diesem Falle sehr nach [ə] oder [ɪ])
>
> ... *durch* between [dʊɐç] and [dʊʁç].
> ([ɐ] in this case has a strong tendency towards [ə] or [ɪ])
>
> (Ulbrich 1972: 93)

More confusing still is the category "r-elision" which should be reserved for those cases in which the phonetics of **r** are no longer deemed to be present. However, elision is also used to cover those cases in which the phonetics of **r** are qualitatively and/or durationally present, but cannot be temporally delimited from the phonetics of the vowel (see Figure 2c-d). Ulbrich's relatively simple classification of the vocalic allophones, then, does not reflect a simpler situation than found for the consonantal allophones, but rather an oversimplification in the description. A potentially more serious problem, however, is the absence of any criteria for establishing whether **r** is phonetically present in an utterance portion. The category "r-indifferent" is used to classify cases where a decision between elision and vocalization could not be made. But the problem is not merely the lack of operational criteria, but has theoretical implications. The lack of any qualitative or durational differences in the vocalic portions of a word pair such as *Bart* ("beard") and *bat* ("offered") is not sufficient grounds for claiming that **r** is not phonetically present in *Bart* in exactly the same way as it is in a close vowel environment such as that illustrated in Figure 2(a).

The formal problem with the generative phonological approach to accounting for phonetic patterns is that its rewrite formalism is too powerful. Coleman (1994) argues that despite repeated attempts at restricting this power, the generative formalism of transformational grammar may still represent no more than an unrestricted rewrite system. The constraint-based approach

of declarative phonology (e.g. Coleman 1994; Scobbie 1993) counters this problem by drastically reducing the mechanisms which can be used to manipulate linguistic structures to unification, which can only combine linguistic structures without changing or removing informational content. In generative phonology the rules derive the phonetics from the phonology by successively modifying and deleting structural information. In a declarative approach the phonological and phonetic levels of abstraction are kept apart and the path between the two is mediated by exponency statements which give the phonological structure a phonetic interpretation. The strict segregation of the phonetic and phonological levels of abstraction and the use of phonetic exponency to mediate between the two levels are of course central features of Firthian prosodic phonology (Firth 1948; Henderson 1949).

Although articulatory phonology (Browman and Goldstein 1989) differs in many respects from declarative phonology, not least because articulatory phonology does not have distinct phonetic and phonological levels of abstraction, it shares one important feature of interest, which is the condition that a gesture cannot be removed or added.

The stark reduction in the manipulative power of both articulatory and declarative phonology not only has formal consequences, but also has implications for the way in which we interpret and account for phonetic data. If linguistic material can no longer be deleted, but the phonetics of a particular phonological object appear to be absent we are forced to consider any one of a number of alternative accounts:

- The phonetics are there, but insufficient attention has been paid to detail.
- The phonetics are there, but are "hidden" behind the phonetics of other objects, and require other recording techniques to make their presence visible.
- The phonetics are there, but are so similar to the phonetics of another object, with which they are cotemporal, that they are not observable regardless of observational detail or recording technique.

All of these have been brought to bear in support of phenomena which generative phonology has dealt with in terms of deletion. Kelly and Local (1989) provide many examples which illustrate that detailed phonetic observation can reveal difference where identity had previously been assumed or where phonetic material which might otherwise have been considered to be absent. X-ray investigation has revealed the presence of lingual activity which is not observable in the acoustic record because it was overlaid by labial closure (Browman and Goldstein 1990). Most controversial of all is the last alternative because the presence of the phonetics is not claimed on the basis of patterns which can be observed whichever method of recording is used. However, there are a number of ways of justifying different phonetic ingredients despite surface identity. Our justification is based on a speaker for whom the vocalic portions in the word pair *Fahrt* ("trip") and *bat* are not observably different[3]:

- The phonetics of **r** in other parts of the *fahr*-paradigm are observable, and can be assumed to be present in *Fahrt* as well.
- In other varieties and for other speakers of the same variety, the phonetics of **r** in *Fahrt* are observable, i.e. the vocalic portions of *Fahrt* and *bat* are different.
- A model which reproduces the observable patterns in words such as *Bier* and *Durst* accounts equally well for the vocalic portion in *Fahrt*.

[3]At present this example is hypothetical as the data base material does not provide suitably comparable material.

Let us summarize the discussion up to this point. We have cast doubt on certain aspects of Ulbrich's descriptive categorization, in particular the treatment of vocalic patterns subsumed under "elision". A generative account of the phonetic patterns has also been rejected on formal grounds.

A declarative, Firthian approach to accounting for the patterns described in the previous section will now be outlined. This involves proposing abstractions at the phonetic and phonological levels of abstraction and deciding which aspects of the phonetic patterns are to be accounted for at which level.

Figure 4 illustrates structural requirements at the phonological level. The different phonetic correlates of **r** are related to different places in the structure of the syllable which in general phonetic terms can be stated as follows:

r at onset:	dorso-uvular stricture of close/open approximation
r at coda:	half-open, central vowel quality

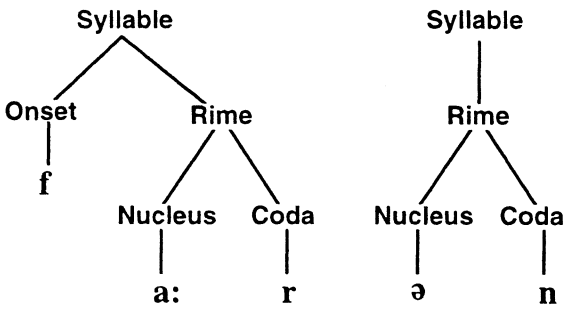
These exponency statements differ little from allophonic statements with the difference that these exponents define the ingredients of utterance and not what is temporally delimitable or directly observable in utterance (cf. Browman and Goldstein's 1992, 'input' and 'output' phonetics). Figure 4 shows that different affiliations of **r** to the syllabic structure can be used to account for tokens of words *fahren* which can exhibit either the consonantal (4b-c) or vocalic (4a) correlates. Indeed, if ambisyllabicity is admitted as a possible structural configuration then tokens of *fahren* with the consonantal correlate can be seen as having **r** at the coda of the first and/or at onset of the second syllable. In the ambisyllabic case (4b) both the vocalic and the consonantal correlates of **r** would be present, in the simple onset case (4c) only the consonantal correlate. In open vowel cases this difference may be difficult to verify, but with other vowel qualities (e.g. *spazieren* "stroll") ambisyllabicity would predict a more diphthongal vocalic portion in the second syllable.

The differences between monophthongal (2c-d) and diphthongal (2a-b) vocalic portions are seen in terms of differences in the amount of temporal overlap between the phonetic correlates of the the vowel and those of **r**, being greater in short than in long quantity syllables. It might be the case – the model presented in the next section implements this – that the greater temporal extent of **r** in short quantity syllables is similar to that found for other consonants following short and long vowels, as has been occasionally reported for other languages (e.g. Nooteboom 1972).

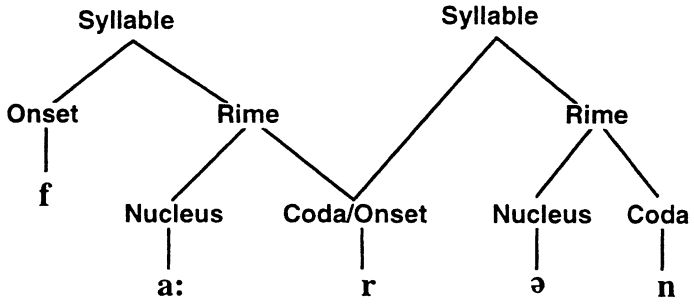
As was clear from the description in the previous section certain aspects of the articulatory and phonatory behaviour associated with **r** are not to be attributed directly to the phonetic correlates of **r**. The presence or absence of vocal fold vibration in certain cases was considered to be a product of the complex interaction between articulatory configuration and air flow/pressure. The absence of voice in voiceless onsets may also in part be due to this interaction, but it is primarily voicelessness as a correlate of such onsets coincident with the dorso-uvular stricture which gives rise to voiceless friction.

In the account proposed in this section, differences in the phonetic patterns associated with **r** are attributable to:

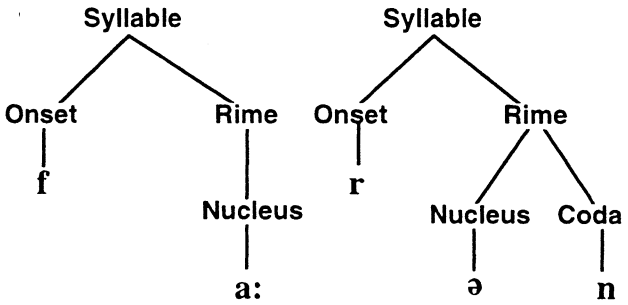
- place in syllable structure (onset, coda);
- extent of temporal overlap of the phonetic correlates of **r** with those of other objects;
- articulatory-aerodynamic behaviour of the vocal tract.



(a) [fɑːn̩]



(b) [fɑːʁən]



(c) [fɑːʁən]

Figure 4: Syllable structures for (a) vocalic and (b) consonantal tokens of the verb *fahren*. If ambisyllabicity is admitted then there are two structural possibilities for consonantal tokens: (b) **r** is at the coda of the first and onset of the second syllable, or (c) **r** is only at the onset of the second syllable.

It is important to consider how this differs from a possible generative account. The necessary phonological abstractions are restricted to syllable structure and the different places in that structure which **r** can take up. Phonetic exponency statements interpret **r** differently depending on its place in structure. The actual patterns which are observed in utterance are the result of the temporal combination of **r**-correlates with those of other objects together with certain articulatory-aerodynamic factors. This allows us to account for a variety of surface phonetics without the need for processes at the phonological level which derive the different patterns from a single base phonetic form.

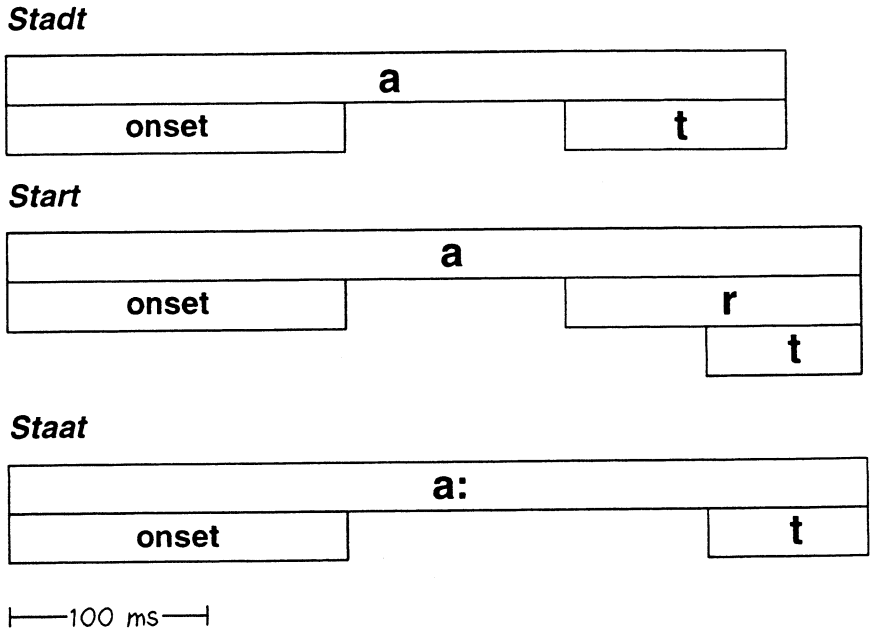


Figure 5: Temporal organization for synthetic versions of the short quantity syllables *Stadt* ("town") and *Start* ("start") and the long quantity syllable *Staat* ("country"). The drawing is to scale.

4 Modelling the phonetics

An interesting and challenging method of verifying the analysis presented in the previous section is to use it to drive a speech synthesizer and thus produce acoustic output. The computational implementation described here has much in common with YorkTalk (Coleman 1992; Local 1992; Local and Ogden 1997) and IPOX (Dirksen and Coleman 1997), both nonsegmental declarative attempts at driving speech synthesizers. This applies to the strict division between phonological structure and phonetic exponency and the way in which phonological structure is given a phonetic interpretation. The model outlined here produces control signals to drive an implementation in C of the Klatt (1980) formant synthesizer.

The phonological structure implemented is essentially that illustrated in Figure 4. The phonetic interpretation of the structure begins by giving each node in the phonological structure a start and end time. The temporal extent of each node encompasses the time span of all daughter nodes. Once temporal information has been assigned, the phonetic exponents of each node are laid down. Timing in the phonetic exponency statements carried out relative to the structural starts and ends, and not in absolute terms. The temporal extent of the phonetic correlates of a particular phonological object are the product of the interaction of the timing assigned to the phonological structure and that encoded in the exponency statements. Differences in the length of consonantal articulations following short and long quantity vowels illustrates one advantage of this separation of temporal information. In long quantity syllables the amount of time assigned to the coda is less than that assigned to the coda of a short quantity syllable. The temporal extent of the phonetic exponents of a coda object in a short quantity syllable is then automatically greater than it is in a long quantity syllable, without this being part of the phonetic exponency statement itself.

Figure 5 shows the temporal make-up of synthetic versions of the monosyllabic words *Stadt* ("town"), *Start* ("start") and *Staat* ("country"). The drawings are to scale and each block represents the time span covered by nodes in the phonological structure of each syllable. Note that

Table 1: Exponency statement to calculate times and values of F1 and F2 for **r** at coda. *start* and *end* refer to the start and end times of the relevant node in the syllable structure; *length* refers to the time span, i.e. $end - start$. *target* refers to the formant value of the vowel. Fx refers to values of F1 or F2.

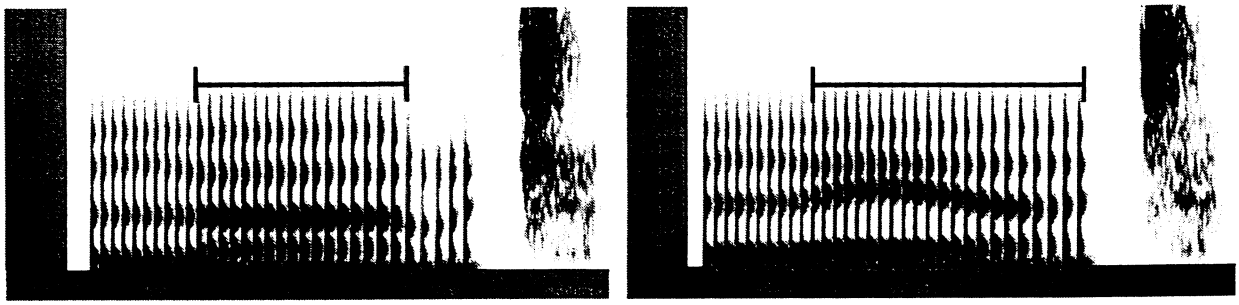
Time [ms]	**Value [Hz]**
$start - length$	$target$
$start$	$Fx + 0.5 \times (target - Fx)$
$end - 20$	$Fx + 0.5 \times (target - Fx)$
$end + 20$	$target$

this is *not* the same as the temporal extent of the phonetic correlates of the objects at each node, which will become clear when we look at part of the exponency for coda-**r** below. The words *Stadt* and *Start* are short quantity, in 5(c) long quantity. The duration of the short quantity syllable in *Start* is greater than *Stadt* due to the increased complexity of the coda. The duration of the coda in both *Stadt* and *Start* is greater than in the long quantity syllable *Staat*. At present the extra duration is assigned to the head of the coda, giving rise to a longer plosive closure in *Stadt* than *Start* and *Staat*. Of greatest interest is the surface similarity exhibited by *Start* and *Staat* despite structural and temporal differences in their make-up. The vocalic portions resulting from each have marginally different durations (the vocalic portion of *Start* is 4 ms shorter).

Part of the phonetic exponency – calculation of times and values for F1 and F2 – for **r** at coda are shown in Table 1. In the left hand column are points in time relative to the structural times (*start*, *end*). So, for instance, for coda **r** $end+20$ refers to a point 20 ms after the end time assigned to the coda node containing **r**. It is now clear how the temporal extent of the phonetic correlates of an object at a particular place in structure differ from the time span assigned to the node itself. The time span of the coda node containing **r** in Figure 5 is 148 ms, but the temporal extent of the phonetic correlates of **r** at this place in structure begin much earlier. The righthand column in Table 1 calculates values for F1 and F2. The qualitative combination of the phonetics of the vowel (*target*) and those of **r** is modelled using a simple locus equation. Values between the calculated points are arrived at using a cosine interpolation. At present, the same locus equation and temporal pattern is used for both F1 and F2 and this undoubtedly represents a simplification, but it is nevertheless sufficient to allow important aspects of the observed patterns to be reproduced.

The separation of timing at different levels allows the different monophthongal and diphthongal qualities to be reproduced using the same exponency statements. The duration of a long quantity syllable is longer than a short quantity syllable (see Figure 5). The duration of the coda in the short quantity syllable is longer than in the long quantity syllable. The temporal extent of the phonetics of **r** at the coda of a short quantity syllable is therefore greater in both absolute and relative terms. In addition to this the time span of the coda node containing **r** is used to define the point in time at which formant movements for **r** begin relative to the start of the node itself ($start - length$). The combination of these factors means that the qualitative combination of vowel and **r** in a short quantity syllable often begins before the end of the onset, as is the case in *Start* in Figure 5.

The consequences of these timing differences are illustrated in Figure 6. The figure shows sonagrams of the short quantity *lernt* ("learns") and long quantity *leert* ("empties"). The phonetics of **r** in *lernt* have already begun during the lateral at the beginning of the syllable, the vocalic portion is monophthongal in quality. In *leert* the diphthongal quality of the vocalic por-



lernt *leert*

Figure 6: Sonagrams of the short quantity *lernt* ("learns") and the long quantity *leert* ("empties"). The horizontal line approximately delimits the vocalic portion in each case.

Table 2: Examples of synthetic short and long quantity **r** and **r**-less syllables. The words *Bart* and *Dirk* have both been synthesized as short and long quantity syllables.

Word (gloss)	**Syllable quantity**
Stadt ("town")	short
Staat ("country")	long
Start ("start")	short
bat ("offered")	long
Bart ("beard")	short
Bart	long
Tier ("animal")	long
Kur ("cure")	long
Kür ("free section")	long
Dirk (proper name)	short
Dirk	long
durch ("through")	short
Storch ("stork")	short

tion is clearly visible in movements of F1 and F2. In this case the phonetics of **r** starts some time after the release of the initial lateral and the phonetics of the long quantity vowel are able to 'peek' through for a short time.

Table 2 contains a list of monosyllabic words illustrating the phonetics of coda **r** with various short and long quantity vowels. The words *Bart* ("beard") and *Dirk* (proper name) have been synthesized as both long and short quantity vowels in an attempt to illustrate one of the areas in which distributional differences between individual speakers and dialects can arise. The **r**-less words *Stadt*, *Staat* and *bat* have been included for comparison. Of particular interest here are comparisons of *Staat* and *Start* as well as *bat* and *Bart*. Surface similarity in the temporal organization of these pairs is now joined by surface similarity in the auditory impression, although the vocalic portions in each pair are acoustically and auditorily different.

5 Concluding remarks

This paper has presented a description of the complex consonantal and vocalic patterns associated with Standard German **r**. The description of the vocalic patterns painted a more complex picture than previous analyses and it was claimed that the descriptive oversimplification may

have arisen from the inability to temporally delineate the vocalic correlates of **r** in many contexts. Section 3 provided an account for these patterns in a declarative, Firthian framework and at the same showed how such an approach ultimately affected the way in which the description itself was carried out because different theoretical assumptions played a part in data interpretation. This was particularly the case in those examples which other analyses had considered to be cases of **r**-elision. Finally, in the previous section a computational implementation of certain aspects of the phonetic and phonological analysis was presented which allowed acoustic and auditory inspection of the analytical claims being made.

The ability to produce very similar phonetic patterns on the basis of significant differences in the phonological structure, temporal organization and phonetic exponency raises interesting questions regarding speech production and perception. As was said above, the phonetic identity in pairs such as *Start* and *Staat* is still hypothetical as it could not be tested on the data base material available. However, if it does prove to be the case that speakers produce word pairs which are acoustically identical, finding out whether the productive mechanisms behind the same surface phonetics, which the model predicts, will be difficult to ascertain, and it is not clear at present how this could be done.

References

Barry, W. J. (1995). Schwa vs. schwa + /r/ in German. *Phonetica 52*, 228–235.

Bickley, C. A. and K. N. Stevens (1986). Effects of a vocal-tract constriction on the glottal source: Experimental and modelling studies. *Journal of Phonetics 14*, 373–382.

Bickley, C. A. and K. N. Stevens (1987). Effects of a vocal-tract constriction on the glottal source: Data from voiced consonants. In T. Baer, C. Sasaki, and K. S. Harris (Eds.), *Laryngeal Function in Phonation and Respiration*, pp. 239–253. Boston: College-Hill.

Bremer, O. (1893). *Deutsche Phonetik*. Leipzig: Breitkopf & Härtel.

Browman, C. P. and L. M. Goldstein (1989). Articulatory gestures as phonological units. *Phonology 6*, 201–251.

Browman, C. P. and L. M. Goldstein (1990). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, pp. 341–376. Cambridge: Cambridge University Press.

Browman, C. P. and L. M. Goldstein (1992). Articulatory phonology: an overview. *Phonetica 49*, 155–180.

Coleman, J. S. (1992). "Synthesis-by-rule" without segments or rewrite rules. In G. Bailly, C. Benoît, and T. R. Sawallis (Eds.), *Talking Machines: Theories, Models and Designs*, pp. 211–224. Amsterdam: Elsevier.

Coleman, J. S. (1994). Polysyllabic words in the YorkTalk synthesis system. In P. A. Keating (Ed.), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, pp. 293–324. Cambridge: Cambridge University Press.

Dirksen, A. and J. S. Coleman (1997). All-prosodic speech synthesis. In J. P. H. v. Santen, *R. W. Sproat, J. P. Olive, and J. Hirschberg (Eds.), Progress in Speech Synthesis, pp.* 91–108. Berlin/Heidelberg/New York/Tokyo: Springer.

Firth, J. R. (1948). Sounds and prosodies. *Transactions of the Philological Society*, 127–152.

Graf, J. and B. Meißner (1996). Neue Untersuchungen zur r-Realisation. *Hallesche Schriften zur Sprechwissenschaft und Phonetik 1*, 68–75.

Hall, T. A. (1993). The phonology of German /ʀ/. *Phonology 10*, 83–105.

Henderson, E. J. A. (1949). Prosodies in siamese. *Asia Minor 1*, 189–215.

IPDS (1994). *The Kiel Corpus of Read Speech*, Volume 1, CD-ROM#1. Kiel: Institut für Phonetik und digitale Sprachverarbeitung.

Kelly, J. and J. K. Local (1989). *Doing Phonology*. Manchester: Manchester University Press.

Klatt, D. H. (1980). Software for a cascade/parallel synthesizer. *Journal of the Acoustical Society of America 67*, 971–995.

Kohler, K. J. (1990). German. *Journal of the International Phonetic Association 20*(1), 48–50.

Kohler, K. J. (1995). *Einführung in die Phonetik des Deutschen* (2 ed.). Berlin: Erich Schmidt.

Kohler, K. J., M. Pätzold, and A. P. Simpson (1995). *From scenario to segment: the controlled elicitation, transcription, segmentation and labelling of spontaneous speech.* AIPUK 29.

Lisker, L. and A. S. Abramson (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word 20*, 384–422.

Local, J. K. (1992). Modeling assimilation in nonsegmental, rule-free synthesis. In G. Docherty and D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, pp. 190–223. Cambridge: Cambridge University Press.

Local, J. K. and R. Ogden (1997). A model of timing for nonsegmental phonological structure. In J. P. H. v. Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (Eds.), *Progress in Speech Synthesis*, pp. 109–121. Berlin/Heidelberg/New York/Tokyo: Springer.

Meinhold, G. (1989). Das problematische [ɐ]. In Slembek (Ed.), *Von Lauten und Leuten. Festschrift für Peter Martens zum 70. Geburtstag (Sprache und Sprechen 21)*, pp. 119–125. Frankfurt am Main: Scriptor Verlag.

Nooteboom, S. G. (1972). *Production and Perception of Vowel Duration.* Ph. D. thesis, Rijksuniversiteit te Utrecht, Utrecht.

Scobbie, J. M. (1993). Constraint violation and conflict from the perspective of declarative phonology. *Canadian Journal of Linguistics 38*, 155–169.

Stevens, K. N. (1987). Interaction between acoustic sources and vocal-tract configurations for consonants. In *Proc. XIth ICPhS*, Volume 3, Tallinn, pp. 385–389.

Ulbrich, H. (1972). *Instrumentalphonetisch-auditive R-Untersuchungen im Deutschen.* Schriften zur Phonetik und Kommunikationsforschung Nr. 13. Berlin: Akademie Verlag.

Viëtor, W. (1894). *Elemente der Phonetik des Deutschen, Englischen und Französischen.* Leipzig: Reisland.

An EPG study of alveolar to velar coarticulation in fast and careful speech: some preliminary observations

Lucy Ellis & W. J. Hardcastle
Department of Speech and Language Sciences, QMC, Edinburgh, UK

1. Introduction

This paper is concerned with the articulatory details of a connected speech process - the assimilation of a word-final alveolar nasal to a following velar plosive, under conditions of varied speech rate.

A commonly observed phenomenon in a number of different languages is the so called 'instability' of alveolar sounds /t, d, n/. These sounds tend to assimilate to following velars and bilabials /k, g, p, b/ - that is, alveolars can take on the place of articulation characteristics of an adjacent segment and either change into or become more like it.

Traditionally, the phonological account of this coarticulatory phenomenon favours a simple binary description reflecting the perceived presence or absence of an alveolar: e.g. *tin can* /tɪn # kan/ → /tɪŋ # kan/. However, studies with electropalatography (EPG), a technique which records spatio-temporal patterns of tongue-palate contact, have established that this description does not capture the full range of patterns that actually occur. Hardcastle and Roach (1979) and Wright and Kerswill (1989) looked at articulation of alveolars in cluster sequences and found that a continuum of patterns across subjects occur ranging from full alveolar closure to 'complete' assimilation (no evidence of tongue-tip/blade contact or lateral contact beyond that which characterises an EPG description of a lexical velar). Varying degrees of 'residual' or 'partial' assimilations occupy the area in between. Hardcastle (1994) found that alveolar nasals are more susceptible to assimilation than plosives. The assimilatory behaviour of alveolar nasals remains a relatively under-researched area.

Nolan (1992) tested the perceptual response of listeners to this continuum of articulations in order to discover what the perceptual correlates of articulatory gradualness might be. He found that while the identification of complete alveolar closure with lexical alveolars is highly reliable, the identification of residual alveolars with lexical alveolars is ambiguous. Furthermore there is no conclusive evidence that listeners are able to recover an alveolar from a 'completely' assimilated alveolar-velar sequence. However, when naive listeners were asked to identify these when presented as a pair with lexical velars, they scored rather better than when presented with them unpaired. If listeners are to some extent able to match a residual phonetic trace of an alveolar stop to their mental lexical representation of that place (and therefore 'restore' it), then it would be reasonable to suggest that speakers too have a more detailed mental lexical representation than is otherwise assumed.

It has long been suggested that differences in lexical phonological form such as 'assimilatory' [ŋ] and lexical [ŋ], will always result in distinct articulatory or acoustic forms, at least for this type of optional assimilation (Nolan, 1992, although the author has since distanced himself from this hypothesis). Reliable evidence for this in EPG studies is not forthcoming although it is possible that studies using alternative instrumentation may establish the existence of a utilisable trace of the alveolar at some level. If not, the prospect

of gestural deletion in this context provides a challenge to the theory of Articulatory Phonology where it is claimed that place assimilation is mainly the result of gestures overlapping and the perception of these as a single gesture. For instance, Browman and Goldstein (1990) consider a casual realisation of the phrase *hundred pounds*. They propose that if the /p/ is superimposed on it, the /d/ gesture is still maintained: 'The bilabial closure gesture may increase its overlap with the preceding alveolar gesture, rendering it effectively inaudible. The overlap of voicing onto the beginning of the bilabial closure yields the [bp] transcription' (p.361). Jun (1996) in his study on place assimilation in *pk* clusters in Korean and English argues that gestural overlap cannot be the sole factor behind place assimilation. He also argues that gestural *reduction* 'is speaker-controlled; it does not result directly from physical constraints on speech-production mechanisms' and therefore is the source of the variability in place assimilation.

Variation in speech rate/style is known to have an effect on assimilation. Other influencing variables include syntactic structure, stress and informational load. Rate has a more significant effect on coarticulation than syntax (Hardcastle, 1985) and a general if not unsurprising finding common to all research in this area is that coarticulation and connected speech processes tend to be applied at fast rather than slow speaking rates. The other tendency is for speaker-specific strategies (different responses to the demands on the articulators of increased rate) to emerge. This dimension of assimilation has only been explored with subject groups as small as two or three - the study reported here is a more systematic, larger scale effort. The robust correlation of rate with connected speech processes is not, however, straightforward. Brown (1990) has demonstrated in an auditory study that assimilation does occur in a 'careful' style of speech. Also, Barry (1985) and Kerswill (1985) using EPG found that while at a faster rate subjects tended to make less alveolar contact, when required to speak fast but 'carefully' this tendency could be overridden.

Durational variations can occur when an articulatory organ does not have sufficient time to complete a given target and so has to 'undershoot' it. This is the basis of Lindblom's 'duration dependent undershoot' model (1963) which was developed from acoustic evidence of vowel reduction. He proposed that articulatory and acoustic undershoot of vowels is a function of *reduction of movement* towards the vowel target due to physiological limitations. This assumes though, that segments in connected speech are reduced equally in duration at fast rates. Another articulatory outcome is an increase in movement velocity (Lindblom, 1990) although this can combine with undershoot (spatial reduction) to give rise to a further strategy (Gay, 1981).

Place assimilation behaviour is not a universal phenomenon. Lindblom (1983) views assimilation as a language specific grammatical rule which represents a categorical change unlike coarticulation which is a continuous motor process. Provisional confirmation of the view that alveolar-velar assimilation in Russian is nowhere near as extensive as it is in other languages, has been provided in an EPG study by Barry (1988). Most interestingly, Farnetani and Bùsa (1994) found that in Italian the alveolar-velar assimilation in /nk/ clusters is always categorical. On the basis of an auditory study of Durham English, Kerswill (1987) notes either an absence or near absence of place assimilation in contexts where it might reliably be predicted to occur in other varieties of English. Once the distribution of place of assimilation behaviour across many languages and accents comes to light we can fully conceive of it as something over which speakers have control. This could have ramifications for theories that are concerned with the precise level of phonetic detail that is specified in speakers' mental representations or phonetic plans.

2. Method

2.1 Stimuli

Speech material was devised so that consonantal combinations would capture potential sites of alveolar assimilation in addition to neutral velar control contexts. These experimental combinations were embedded in the sentences: *"I can't believe the ban cuts no ice "* /n#k/ and *"I've heard the bang comes as a big surprise"* /ŋ#k/. Further material was devised to capture two other consonantal combinations: /n#t/ *"I'm not surprised the ban touched a raw nerve"* and /ŋ#t/ *"I reckon the bang toughened her resolve"* the results from which are not reported here.

The vocalic environment was kept as consistent as possible. Bordering vowels were /a/ & /ʌ/ and the /a/ vowel was preceded by a bilabial stop to eliminate the possibility of any lingual coarticulatory effects on the target consonants. This latter control will be particularly advantageous in the light of a planned EMA experiment using similar test stimuli where the only source of coarticulation on tongue trajectories will be the tongue tip and dorsum themselves. Low vowels were used to flank the consonants because vocalic tongue-palate contact (characteristic of high front vowels for instance) would be minimal. Also it was predicted that a rather 'front' velar occlusion would be achieved after /a/, more easily observed on EPG printouts. Voiceless plosives were selected to follow the nasal in all experimental sentences so that the offset of voicing for the nasal could be directly measured in relation to the transition from nasal stop to oral stop. This is especially useful in those cases where phonetic differences are sought between 'assimilatory' /ŋ/ to /k/ and lexical /ŋ/ to /k/, *...ban cuts.../...bang comes...*, even though these are not strictly minimal pairs.

A further 4 'filler' sentences of no experimental interest were added to the original 4 to distract the speakers from the presence of near minimal pairs. 10 repetitions of each test item were produced.

2.2 Speakers

10 speakers with EPG palates were recorded. All but 2 of the subjects were female and overall the subject group represented a fairly wide range of regional accents. 4 of them spoke what might be called Standard Southern British, one was Australian, one was Northern Irish, one spoke a variety of North Eastern English and the remaining 3 spoke Scottish English from 2 different regions.

2.3 Data collection

The technique of electropalatography (EPG) was used to record the timing and the location of tongue contact with the hard palate during continuous speech. During the course of the recording each speaker wore an artificial palate embedded with 62 silver electrodes which are activated when contact occurs. The details of contact are then stored on computer. The palate can be roughly divided into three zones: alveolar region (rows 1 and 2), palatal region (rows 3-5) and velar region (6-8). Although palates are custom made for speakers these regions follow predetermined anatomical landmarks which target phonetically significant areas. The sampling rate of this system is 100 frames per second and the acoustic signal is 10kHz.

Before the experiment speakers wore their palates for an hour to adjust, although they in fact required little acclimatization. Most of the subjects were experienced wearers of EPG palates and had been involved in other experiments using the technique.

The experiment fell into two parts. The aim of the first part was to elicit careful speech and all subjects were instructed to read each sentence *slowly and clearly*. No particular instruction was given with respect to prosodic phrasing. The aim of the second part was to elicit fast/casual speech although genuinely casual speech is notoriously difficult to acquire under laboratory conditions. The material for this part was identical to the first but the 80 sentences were arranged in groups of 3 and filler sentences were purposely distributed to avoid 'clustering' of experimental sentences. Subjects were instructed to read out each group of sentences in a *rapid and casual style* one after the other avoiding obvious pauses in between each sentence. A time limit of 5 seconds was imposed on the delivery of each group of sentences and time was allowed for subjects to practice both the fast groups and the individual careful sentences. The time constraint automatically ruled out any attempt on the part of speakers to impose too complex an intonational structure on each sentence. All subjects perceived this time limit as a challenge so it was an effective measure against over-awareness of the test items. The sentences in the first half of the experiment and the groups in the second were individually cued during recording with a pause between each.

2.4 Data analysis

Electropalatographic measurements were taken from the EPG trace and acoustic measurements were taken from the waveform/spectrogram. For all experimental sequences annotation points were made from the onset of the vowel before the word boundary (/a/ in all cases) up to onset of the vowel after the word boundary (/ʌ/ in all cases).

5 annotation points were made for sequences where a single place of articulation was achieved i.e. /ŋ#k/ ...*bang comes*... in fast and careful rate and /n#k/ ...*ban cuts*... in the fast rate, where there was a complete absence of alveolar contact or it was insufficient to justify annotation. It was also predicted that, for some speakers, a careful articulation of this latter sequence would motivate an assimilation of this sort. 7 annotation points were made for sequences where there were 2 clearly observable adjacent places of articulation. That is, in the context /n#k/ in the careful and fast speech rate where the alveolar target was achieved. A database was created which stores the EPG frame corresponding to each annotation point and its time value.

Figure 1 below shows a waveform, spectrogram and EPG display of one canonical repetition of ...*ban cuts*...in the careful speech condition. The numbered annotation points are indicated on the spectrogram and the timing of these in relation to tongue-palate contact during the sequence is marked on the EPG display below (in each frame of EPG data the top of the schematic plan is the alveolar ridge and the bottom row approximately corresponds to the junction between the soft and the hard palate). These annotation points are defined in Table 1 below.

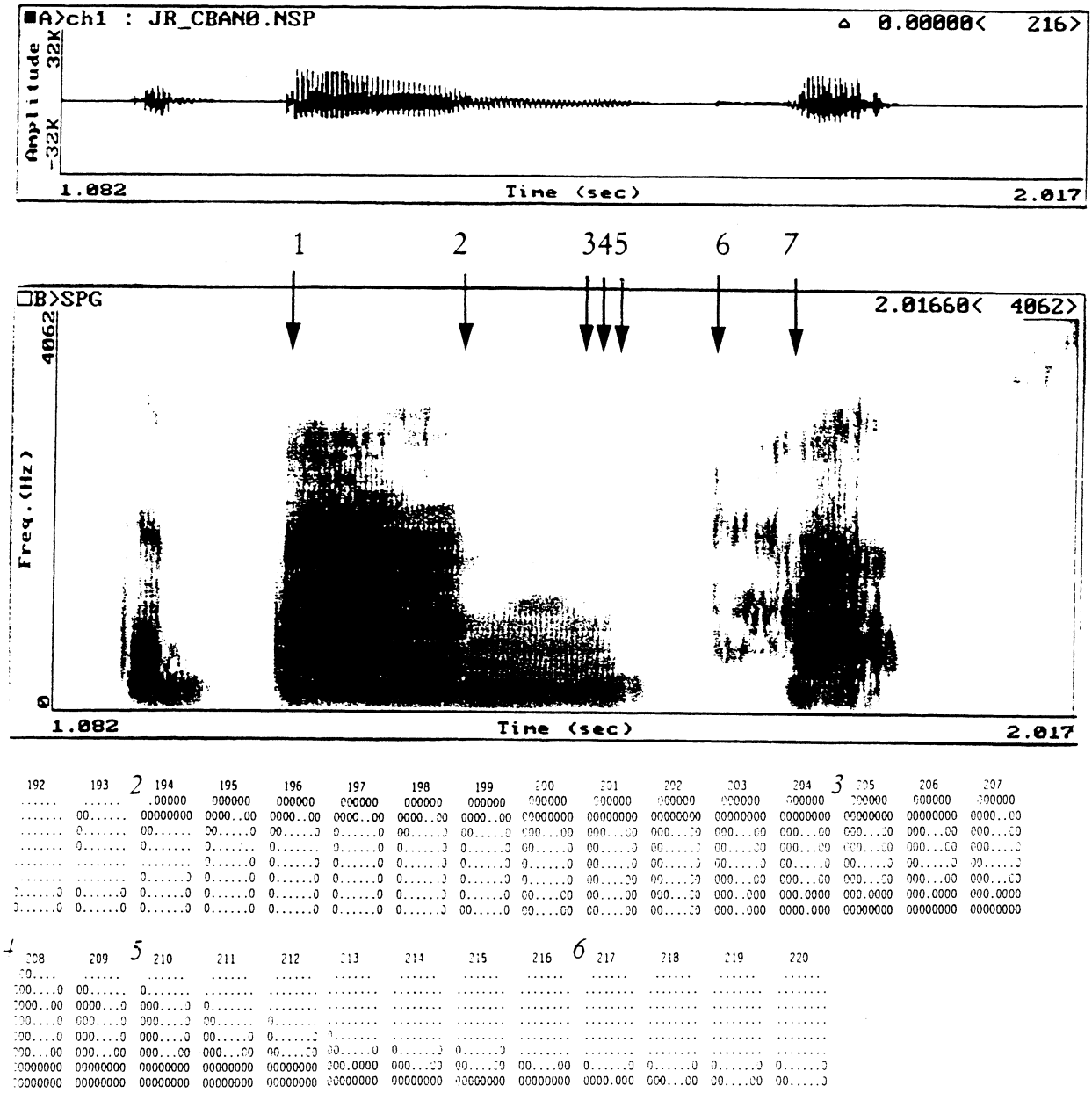


Figure 1: waveform, spectrogram and EPG display of...*ban cuts*... careful speech (subject JR). Numbered arrows represent acoustic and EPG-defined annotation points.

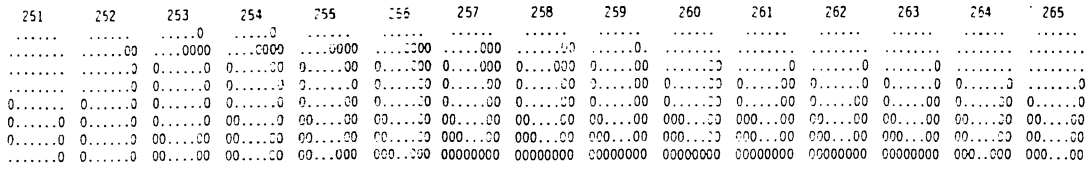
TABLE 1. Definition of annotation points: acoustic and EPG

no.	type	definition
1	acoustic	Onset of periodicity for the vowel /a/
2	EPG	Onset of mid sagittal contact in first 3 rows for alveolar /n/
3	EPG	Onset of complete or maximum constriction in row 8 for velar /k/
4	EPG	Earliest appearance of loss of contact for release of /n/
5	acoustic	End of nasal formant structure for /n/
6	EPG	Earliest appearance of loss of constriction for velar /k/
7	acoustic	Onset of periodicity for the vowel /ʌ/

N.B. The order of annotation points as they appear in the example in Fig.1 & Table 1 is, obviously, subject to variation. For example, formant structure for /n/ can end before the alveolar closure is released.

For analysis of the /n#k/ data a clear distinction had to be made between a 'canonical' alveolar or allophone of an alveolar and an assimilation of an alveolar. For a preliminary indication of the assimilatory trends, partial assimilations were subsumed into the general category of assimilation. To be classed as an alveolar, a pattern had to show mid sagittal contact in the first three rows of the EPG palate. Thus according to this definition, in Figure 2 (a) we see an allophone of /n/ while in (b) we see a partial alveolar articulation which here would be labelled an allophone of /ŋ/.

2(a)



2(b)

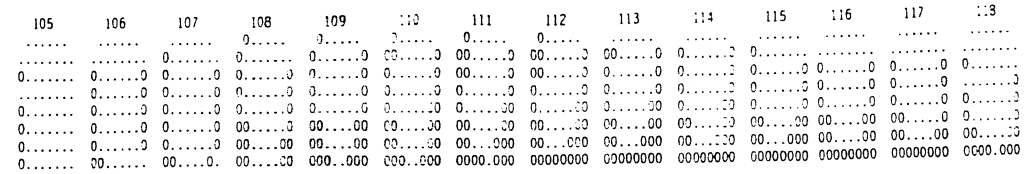
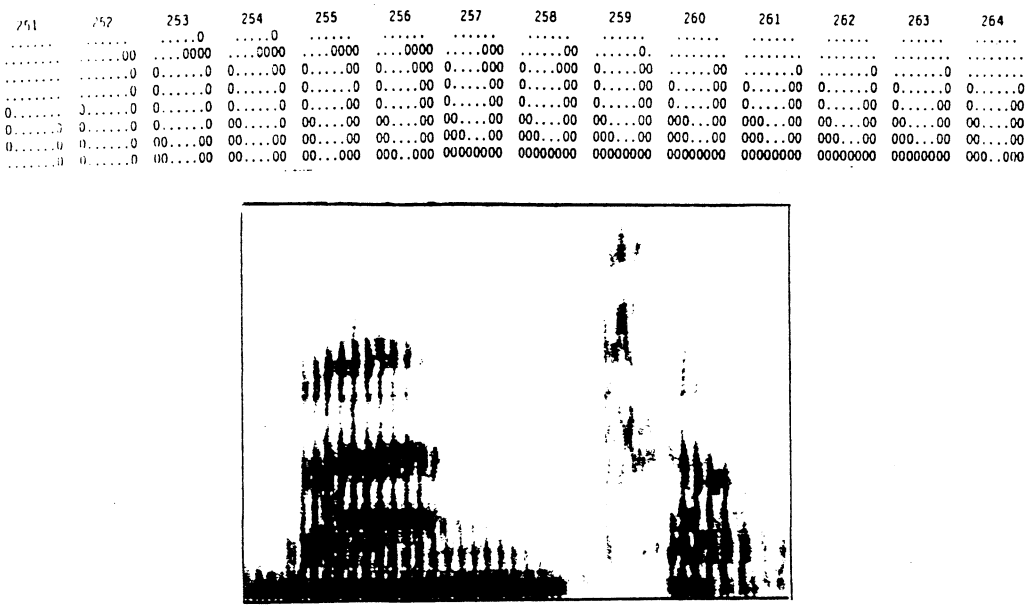


Figure 2 EPG patterns for...*ban cuts*... fast speech 2(a) annotation of an alveolar stop (JSC) 2(b) no annotation of an alveolar stop (WCM)

Although Fig.2 (b) is not classed as an alveolar articulation, the tongue has made the supporting lateral gesture for /n/ and as a result of tongue-tip raising, an acoustic correlate similar to one of a full alveolar might be expected. Figure 3 shows spectrograms for (a) a partial closure at the alveolar ridge and (b) a complete assimilation for *...ban cuts...* fast rate. F2 in (b) stays level while in (a) it drops slightly as the tongue tip makes contact. Notice also the lack of transition in (b) for the velar constriction. More thorough acoustic analysis of complete assimilations in this context and their comparison with lexical velars is planned.

3(a)



3(b)

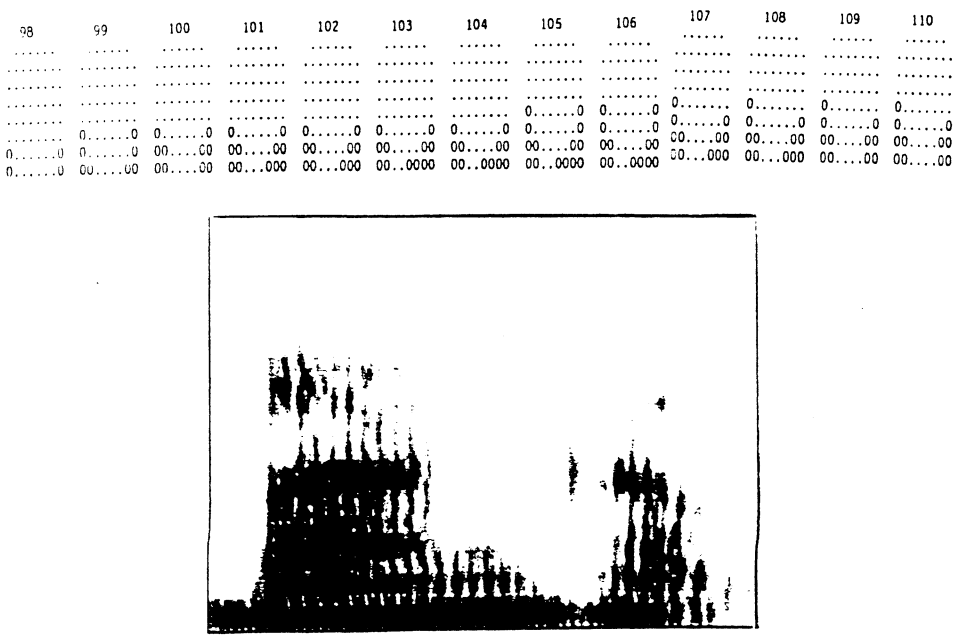


Figure 3 Spectrograms and EPG patterns for (a) incomplete alveolar closure (b) assimilated alveolar. Subject JSC

3. Preliminary Results and Discussion

The distribution of assimilations for /n#k/ in both fast and careful speech for all subjects are shown in Table 2.

TABLE 2

	careful speech	*fast speech*
assimilated	**3**	**56**
non-assimilated	**97**	**44**

The first thing to notice is that there are surprisingly few assimilated alveolars in the careful speech condition. These were produced by 2 subjects JF and SW who rather appropriately happened to be subjects who produced 100% assimilations of the same sequence in the fast rate. This might suggest that their systems are more predisposed to connected speech processes than others. The other thing to note from Table 2 is that although there are more assimilations in the fast rate as might be predicted, they by no means dominate the picture.

A breakdown of the results for fast speech /n#k/ for each subject is shown in Figure 4 below. There is considerable gross variation between speakers with regard to occurrence of assimilation. 2 subjects never assimilated in fast speech and 3 subjects always did (one subject assimilated for 9 out of 10 repetitions). The picture becomes more complex, however, if the patterns for the subjects in the middle of the graph are considered in more detail. Subjects FG, WCM, JSC, TH sometimes 'assimilated' and sometimes did not. But the graph does not tell the whole story, of course, since partial assimilations are not represented here. While the values for subjects JSC and TH appear the same on the graph they do in fact use entirely different reduction strategies as interpreted from the EPG contact patterns.

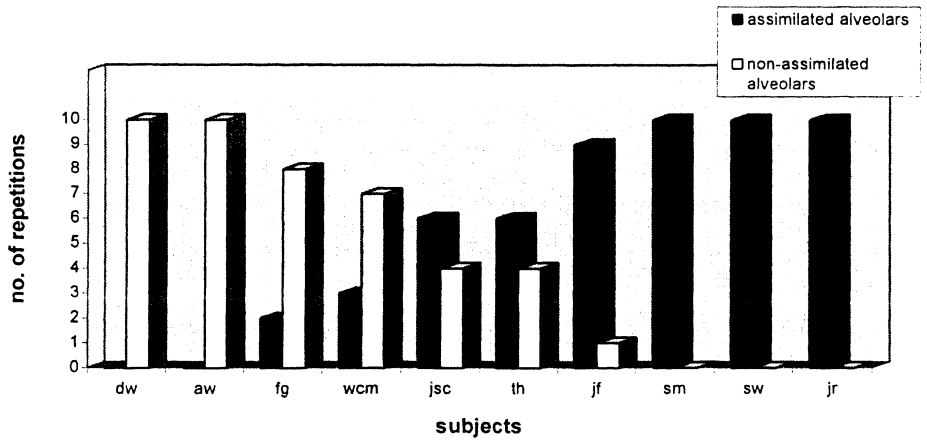


Figure 4 Distribution of assimilations for all 10 subjects ...*ban cuts*... fast speech

The EPG patterns for TH and FG suggest the adoption of a binary segmental strategy. That is, either full alveolar contact is achieved for target /n#k/ or there is segmental substitution with something that looks very similar to /ŋ/. The latter appears to be a restructuring option that precludes gradient partial alveolar assimilations. But partial assimilations *are* tolerated by JSC and WCM. Figure 2 shows the type of 'undershoot' articulations made by these subjects.

On the basis of this observation it would appear that not all speakers have the same 'allophone tolerance' for /n/ in this phonetic environment and in the articulatory domain at least. TH and FG behave here as speakers whose mental representation of /n/ specifies only a single alternative articulatory target. JSC and WCM are, however, speakers for whom permissible realisations are not categorical but include presumably unlimited graded intermediate articulations. One wonders how the contents of a mental lexical representation for these speakers might be expressed. It was thought possible that the discreteness of the two groups in this respect is a function of one of the groups speaking at a faster rate than the other. Target undershoot as a possible manifestation of inertial effects on the articulators is probably more closely associated with fast speech rate than segmental substitution, since the latter is known to occur even at careful rate. Measurements of duration of the sentences did not show JSC's and WCM's fast speech rate to be significantly faster than that of TH and FG (see Figure 5 below). Furthermore, the variability range for the fast speech repetitions for one group is not broader or narrower than for the other.

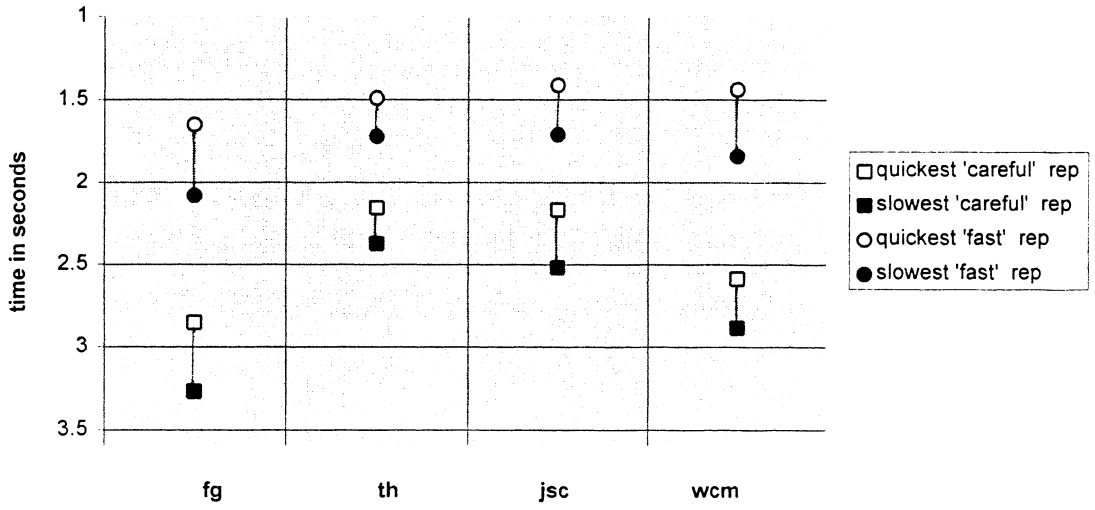


Figure 5. Duration ranges of careful and fast speech sentences for 4 subjects.
N.B. quickest utterance times at the top end of the graph.

Residual alveolars of the type that showed up in the data for JSC and WCM were almost completely absent in the data for the other 8 subjects. JSC and WCM were the only speakers whose alveolar-velar articulations were clearly gradient. They are, incidentally, both from West Scotland.

Figure 6 shows the EPG patterns for all 10 repetitions of ...*ban* cuts...fast rate, for subject JSC. Each line shows a single repetition and captures the realisation of /n#k/. They are ordered to show a gradation from full alveolar contact to complete assimilation via consonant shortening of /n/ and target undershoot.

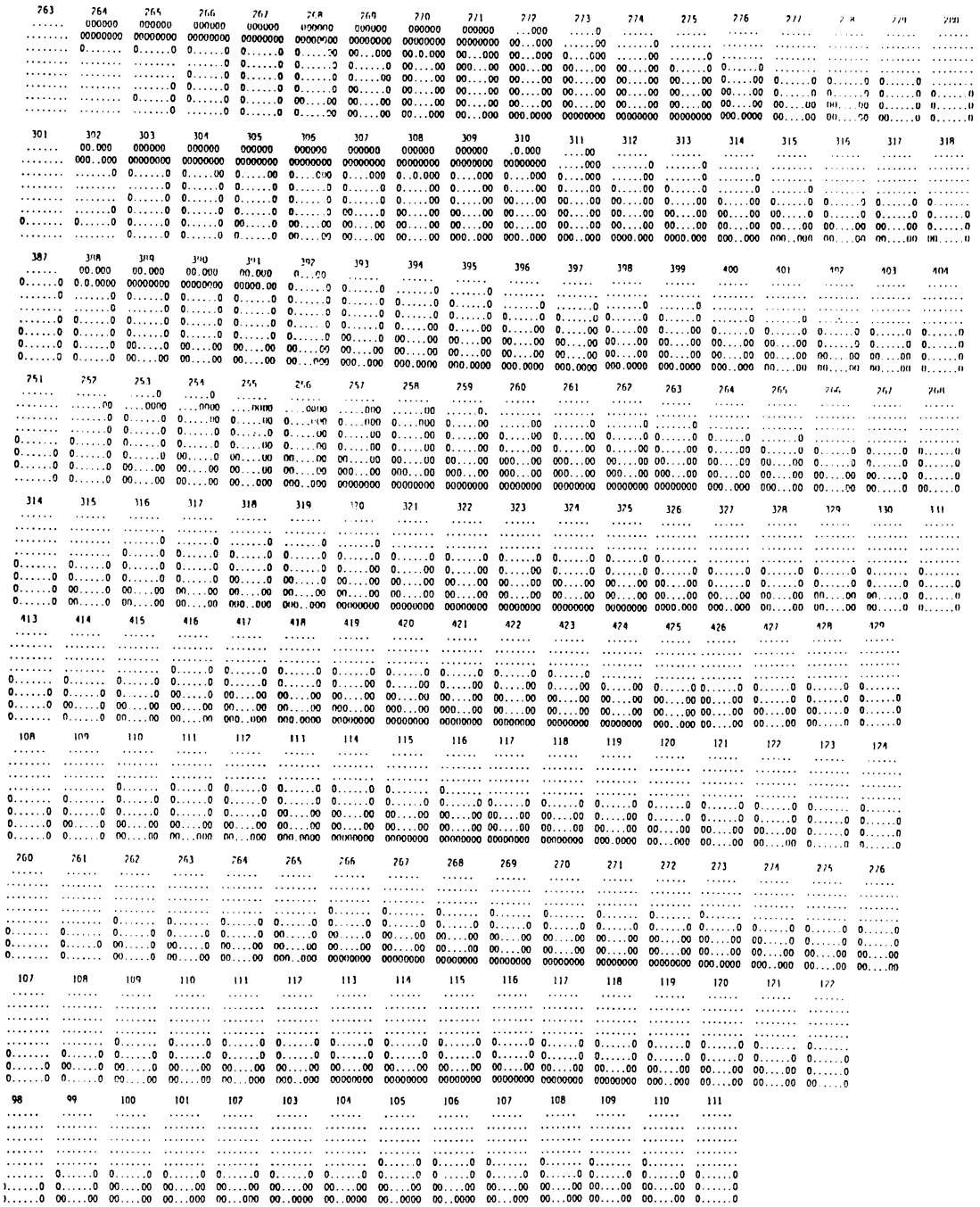


Figure 6 EPG patterns for all 10 realisations of /n#k/ sequence in ...*ban cuts*...fast speech. (JSC)

The EPG contact 'totals' profiles for the patterns shown in Figure 6 are plotted in Figure 7 (a) below. Contact totals for all 10 repetitions from the beginning of the vowel /a/ are superimposed on a single graph. For each repetition there are two curves representing the amount of electrodes contacted on the palate frame by frame in the alveolar region (front 3 rows) and the velar region (back 3 rows). 7 (b) shows 10 realisations of the sequence...*ban cuts*...for fast speech by subject AW. 7(c) shows 10 realisations of the same sequence by AW but for careful speech.

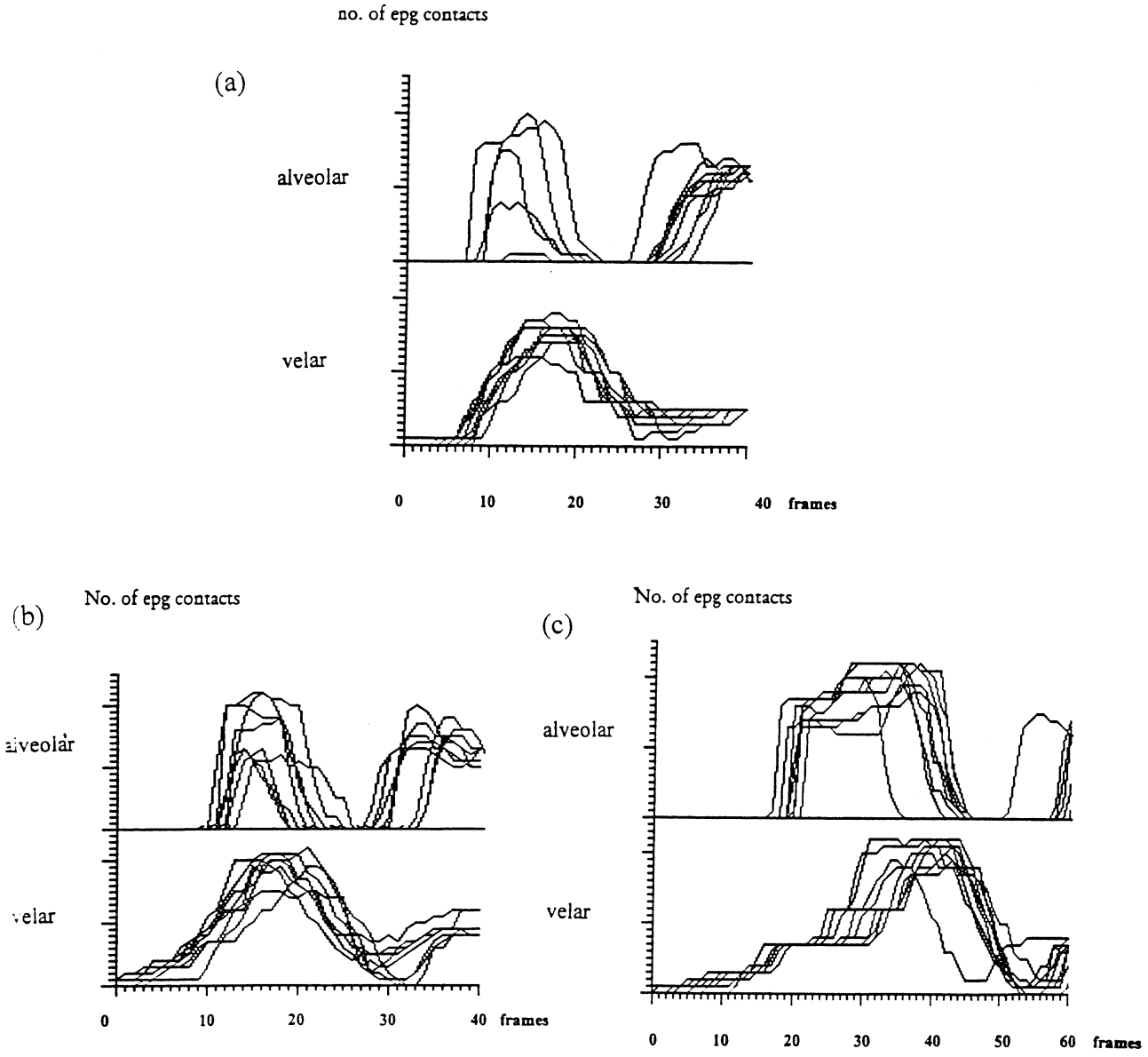


Figure 7 EPG 'totals' graphs showing tongue-palate contact in both alveolar and velar regions for sequence /n#k/ in ...*ban cuts*...The x-axis represents time in frames. 7(a) fast speech, subject JSC; 7(b) fast speech, subject AW; 7(c) careful speech, subject AW

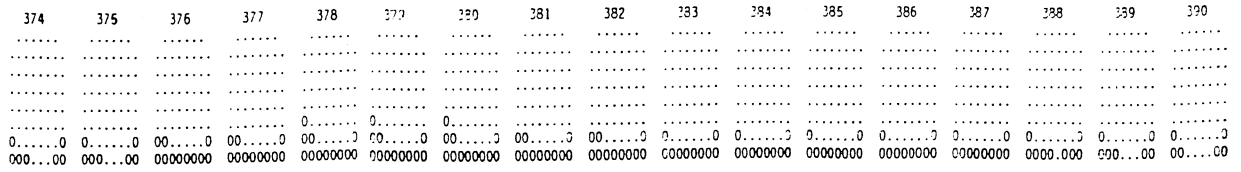
In 7(a), the curves describing residual alveolar articulations are plotted. The shallowest curve of all represents repetition number 5 on Fig. 6 where for 5 EPG frames one electrode was contacted in the third row. The next shallowest curve is the fourth repetition on Fig. 6 where only a partial occlusion is formed and with a slightly more gradual onset than reps 1-3. Of course only 5 'alveolar' curves are shown on this graph because there was no alveolar gesture discernible on the EPG trace for the other 5 repetitions.

AW was one of only two subjects who produced no assimilations for /n#k/ in either careful or fast speech. This subject was quite variable in the fast rate, however, with respect to timing and especially to amount of total contact for /n/ (b). The main reduction strategy adopted was consonant shortening which was sometimes combined with reduction of alveolar contact. This combination produced the variability in alveolar curves in (b), an effect less apparent in (c) for the careful rate. Notice that compared to JSC, the alveolar gesture in (b) is initiated later. Also, the timing relationship between maximum closure for the alveolars and velars for AW in (b) is not the same as that for JSC in (a). AW frequently produced double articulations in careful speech and so maximum contact for each region was often simultaneous. JSC, however, tended to avoid double articulations as illustrated in Fig 6. Kinematics of tongue tip/blade and dorsum are maximally contrasted in 7(c) regarding time interval between initiation of gesture and maximum tongue-palate contact for that gesture. Here for AW velar contact builds up during the markedly less gradual tongue-tip articulation for each repetition. Timing of the release of the alveolar and velar in (c) is likewise consistent apart from one repetition. Of course, in the careful rate (c) the whole sequence up to the complete loss of contact for the /k/ takes around 50 frames whereas for fast speech (a) it takes around 30 frames.

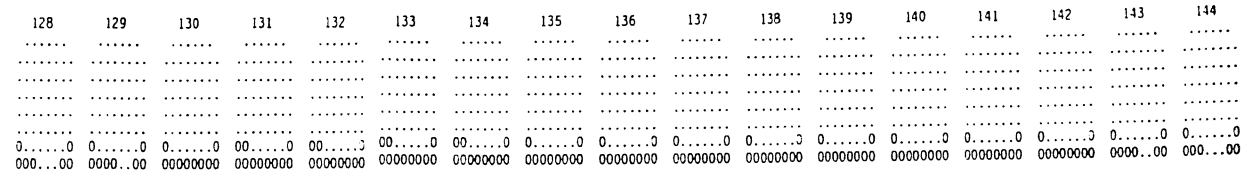
A quite different reduction strategy altogether is used by 3 subjects occupying the right-hand side of the graph in Figure 4 - SW, SM, JR - who all speak Standard Southern British. For all their /n#k/ fast sequences there was a complete absence of alveolar closure and little evidence of lateral contact further forward than that characterising a lexical velar. If lateral contact was in evidence, then for the token to be classified as a residual alveolar according to the definition adopted in this paper, there would be contact along the sides of the palate least one row further forward than for a lexical velar produced by the same speaker under the same speech rate condition. When to compared to the EPG patterns for these subjects' fast all-velar sequences ...*bang comes*..., the 'assimilated' /n/ to /k/ patterns are very similar. It would appear that /n/ has been deleted in this context.

For one subject in particular there was even less EPG contact in the velar region for the 'assimilatory' velars than for lexical velars. In other words the habitual posterior tongue placement is more retracted. Some examples are shown in Figure 8 below.

(a)

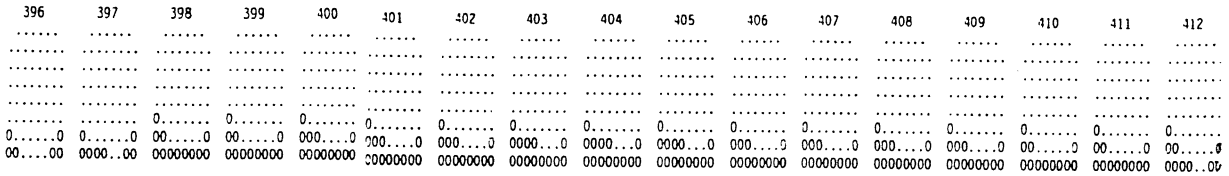


(b)

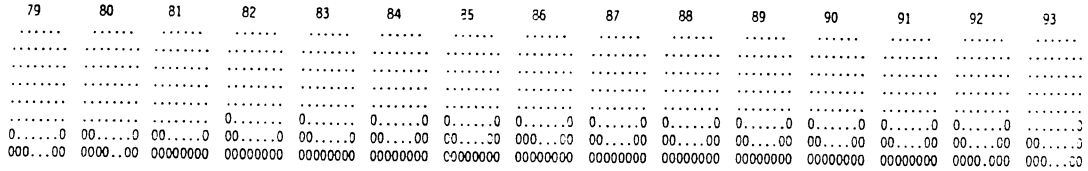


...ban cuts...

(c)



(d)



...bang comes...

Figure 8 fast speech EPG patterns showing (a) & (b) velar tongue position for complete assimilation in 2 repetitions of...*ban cuts*.... contrasted with tongue position for 2 repetitions of all-velar control sequence in...*bang comes*...(c) & (d). Subject JR

There has been some speculation that this retracted assimilatory velar pattern, far from being an extreme form of complete assimilation, is residual evidence of a tongue body configuration appropriate for tongue tip elevation.:

> ...as the tongue tip moves up towards the alveolar ridge, the blade and pre-dorsum become concave, which reduces the amount of lateral contact in the pre-velar area. At the same time, this tongue shape will cause the velar contact itself to be more retracted.
>
> (Wright and Kerswill, 1989)

If this was the case, place assimilation here could be accounted for by Articulatory Phonology where gestures can reduce in magnitude and overlap, often to the extent that place of articulation is lost, although segments are never 'deleted'. If so-called complete alveolar assimilation is not necessarily incompatible with tongue tip elevation then the next question is whether the tongue tip is raised for some or all of the habitual complete assimilations illustrated in Fig 8 (a) & (b). But for those residual articulations produced by JSC and WCM where the tongue is making quite advanced contact with the sides of the teeth leaving less area of the tongue-tip /blade to manoeuvre, it might be the case that tongue-tip elevation is somewhat inhibited. But since the tongue-tip can function as a semi-independent articulator in relation to the tongue dorsum, it is possible that the tongue-tip is still able to describe a substantial raising (and looping?) trajectory although not of the same magnitude as that which, it is suggested, may give rise to the type of retracted velar seen in Fig 8 (a) & (b).

4. Conclusion

The two most interesting questions raised from this study have resulted from the identification from this data set of four broad inter-speaker reduction strategies (namely, one which reduces the consonant but avoids assimilation; one which avoids non-assimilation, one which involves binary variation and one which involves non-binary variation). Firstly, why does the distribution of residual alveolar gestures for this data set concentrate around only 2 subjects? While JSC and WCM produce full alveolars and 'complete' assimilations *aswell*, for subjects TH and FG the production of either of these is the result of a categorical binary option, intermediary articulations being somehow 'blocked'. To what extent can this distinction between these two groups of speakers be attributed to the absence or presence of something in their mental representation?

Electropalatography is a technique which can provide useful information about tongue-palate contact during connected speech leading us to the formation of important research questions. It cannot, however, provide information on proximity of the tongue to passive articulators and does not give an indication of which part of the tongue is involved in contact. In order to confirm or challenge the limited interpretations of the EPG data presented here, a combined EMA/EPG experiment using similar stimuli is planned involving a subject from each of the two groups. EMA tracks the movement, in the x-y plane, of sensors attached to the mid-line of the tongue and provides complementary information to EPG on tongue-dynamics and the overall configuration of the tongue during connected speech. The complete assimilations produced by the 'binary' subjects and those produced by the 'non-binary' subjects will hopefully be distinct in terms of tongue-tip elevation. The hypothesis is that the complete assimilations of the former group will involve less or no tongue-tip elevation compared to the latter, since for the latter the EPG evidence suggests that contact with the alveolar ridge is still being targeted. Clear, if reduced coronal trajectories, falling short of contact could be in evidence some of the time or on a more consistent basis. Kühnert (1993) in her study of the assimilatory behaviour of voiceless plosives reported that in cases of EPG-defined complete assimilation, clear coronal elevations are sometimes present and sometimes are not. Speaker-specific reduction strategies as reported here could prove to be the cause of this variability. The second question which could be answered by EMA observations of tongue-tip movement during assimilation, evolves from the type of data shown in Fig 8 (a) & (b).

These speakers are either consistently preserving the raising gesture part of the alveolar target, which could explain the retracted velar position and lack of lateral contact, or they are producing an extreme form of assimilation, or possibly alternating these articulations. The identification of habitual articulations from EPG data may not be mirrored in EMA data for the same sequences.

References

Barry, M. (1985) A palatographic study of connected speech processes. *Cambridge Papers in Phonetics and Experimental Linguistics* 4

Barry, M. (1988) Assimilation in English and Russian. Paper presented at the colloquium of the British Association of Academic Phoneticians, Trinity College, Dublin, March 1988

Browman, C.P. and Goldstein, L.M. (1990) Tiers in articulatory phonology, with some implications for casual speech. In Kingston, J. & Beckman, M.E. (eds.) *Papers in Laboratory Phonology I* Cambridge University Press, 341-76

Brown, G. (1990) *Listening to Spoken English.* Longman, Harlow 2nd edition

Farnetani, E. and Bùsa, M.G. (1994) Italian clusters in continuous speech. *Proceedings of 1994 International Conference on Spoken Language Processing* 1, 359-62 Yokohama, 18-22 September

Gay, T.J. (1981) Mechanisms of the control of speech rate. *Phonetica* 38, 148-58

Hardcastle, W.J. (1994) Assimilation of alveolar stops and nasals in connected speech. In J. Windsor Lewis (ed.) *Studies in General and English Phonetics in Honour of Professor J.D. O'Connor.* Routledge, 49-67

Hardcastle, W.J. (1985) Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication* 4, 247-63

Hardcastle, W.J. and Roach, P.J. (1979) An instrumental investigation of coarticulation in stop consonant sequences. In P. Hollien and H. Hollien (eds.) *Current Issues in the Phonetic Sciences* Amsterdam: John Benjamins A.G. 531-540

Jun, J. (1996) Place assimilation is not the result of gestural overlap: evidence from Korean and English. *Phonology* Cambridge University Press 13, 377-407

Kerswill, S. (1985) A sociophonetic study of connected speech processes in Cambridge English: an outline and some results. *Cambridge papers in Phonetics and experimental Linguistics* 4

Kerswill, S. (1987) Levels of linguistic variation in Durham. *Journal of Linguistics* 23, 23-49

Kühnert, B. (1993) Some kinematic aspects of alveolar to velar assimilations. *Forschungsberichte des Institut fur Phonetic und Sprachliche Kommunikation der Univeritat Munchen* 31, 263-72

Lindblom, B. (1963) Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America* 35, 1773-81

Lindblom, B. (1983) Economy of speech gestures. In P.F. McNeilage (ed.) *The Production of Speech* New York: Springer Verlag 217-45

Lindblom, B. (1990) Explaining phonetic variation: a sketch of the H and H theory. In Hardcastle, W.J. and Marchal, A. (eds.) *Speech Production and Speech Modelling* Dordrecht: Kluwer

Nolan, F. (1992) The descriptive role of segments: evidence from assimilation. In D.R. Ladd and G.J. Docherty (eds.) *Papers in Laboratory Phonology II* Cambridge University Press, 261-80

Wright, S. and Kerswill, P. (1989) Electropalatography in the study of connected speech processes. *Clinical Linguistics and Phonetics* 3, 49-57

The status of external sandhi in Cree

Kevin Russell*
University of Manitoba

1. Introduction

The Cree language has a set of vowel deletion and coalescence processes that apply at word boundaries. These processes seem to be perfect examples of the kind of post-lexical phonological rule which has led some researchers to propose complicated extensions to the theory of phonology. In this paper, I argue that Cree sandhi is not a true phonological rule, but is the kind of gradient phenomenon which can be handled by a number of existing theories of phonetics and the phonetics-phonology interface. As Cree sandhi joins the growing list of apparent phonological rules which are better assigned to the phonetic domain, it lends support to the hypothesis that phonology is categorical and lexical, and that post-lexical phonology need not exist.

1.1 Cree background

Cree (an Algonquian language spoken in much of Canada) has an essentially triangular vowel system with length distinction. There are three short vowels, conventionally written *i*, *a*, and *o*, though *o* in fact ranges between mid and high. Each short vowel has a long counterpart, conventionally written with a circumflex accent. There is in addition a long *ê*.

(2) Cree vowel system

short	i		o
		a	
long	î		ô
		ê	
		â	

Cree has optional processes that affect vowels that meet across a phonological word boundary. Although it is the phonological reality of these processes that is at issue in this paper, I will describe them in this section as if they were absolute and unquestionable, without the liberal sprinkling of the word "apparently" that strict accuracy would call for.

In the most general case, a short word-final vowel can be deleted. The following word-initial vowel, if it is short, will undergo compensatory lengthening. As a special case, word-final *a* and word-initial *i* will coalesce into long *ê*, rather than the expected *î*. Wolfart's (1973) initial description of these processes is given in (3).

---

* I would like to thank the presenters and audience at the Conference on the Word as a Phonetic Unit for their useful and stimulating discussion. I am indebted to H.C. Wolfart for long-standing discussions on all matters Cree and, especially for this paper, allowing me access to the tapes of Mrs. Minde's autobiographical narratives. This research has been conducted under grant 410-94-1608 from the Social Sciences and Humanities Research Council of Canada, whose support is gratefully acknowledged.

(3) $\breve{V}_1 + V_2 \rightarrow \bar{V}_2$
a + i → ê

As another special case, a word-final short o will delete and cause compensatory lengthening on the following vowel, but will continue to be realized as labialization on the preceding consonant.

Some examples from Minde (1987) are given in (4-6). The first line gives an idealized representation of the sentence, giving each word in the canonical form it has as the output of lexical phonology. The second line gives the actual sentence, as transcribed by the editors, showing the effects of sandhi.[1]

(4) êwak**o** anim**a** êkos**i** ê - kî - isi - pimâcihocik kayâs ayisiyiniwak
êwak**w** ânim êkos ê - kî - isi - pimâcihocik kayâs ayisiyiniwak
'That is how that people made a living long ago.'

This paper will concentrate on two particular kinds of sandhi: i) the result of short *a* and long *ê* meeting across a word boundary, as illustrated in (5), and ii) the result of the complementizer particle *kâ* coalescing with a following short *i*, as illustrated in (6).

(5) napêwasikana mân**a** ê - kî - osîhâcik
napêwasikana mân ê - kî - osîhâcik
'They used to make men's socks.'

(6) ê-kî-âpacihtât mâna ânima âya, 'astinwân' **kâ-i**siyîhkâtêk
ê-kî-âpacihtât mân ânim âya, 'astinwân' **k-ê**siyîhkâtêk
'...she used to use "sinew", as it is called.'

1.2 Theoretical implications

A significant amount of work has gone into making phonological theory capable of dealing with between-word processes of the kind illustrated by Cree external sandhi. Kaisse (1985), for example, argues for a typology of postlexical rules, distinguishing between P1 rules (which crucially need to refer to morphological or syntactic properties or constituents) and P2 rules (the more familiar kind of postlexical rules which apply without regard for syntax). There has been an understandable reluctance to accept the theoretical consequences of P1 rules. Condoravdi (1990) reanalyzes sandhi rules in Greek, one of Kaisse's central examples, so that they no longer need to refer to syntactic consituency but simply to constituents of the prosodic hierarchy (though crucially a prosodic hierarchy with a level between that of the word and phrase). Hayes (1990) offers a proposal for "pre-compiled phrasal phonology", in which the effects of phrasal rules can be computed entirely within the lexical component, leaving the postlexical component just the job of choosing the appropriate lexically derived representation for each word.

A more radical approach to the problems posed by sandhi rules is to question whether

---

[1] Word-by-word glosses for these examples are:
(4) aforementioned that:one in:that:way COMP - PAST - thus - they:made:a:living long:ago people
(5) men's:socks HABIT COMP - PAST - they:made
(6) COMP - PAST she:used HABIT that the:one, sinew COMP - it's:called

the phenomena really exist or, more precisely, whether the phenomena are really better handled by the categorical devices of phonology rather than by the independently needed gradient devices of phonetics. Kaisse (1985: 115, fn 5-6) acknowledges that Greek sandhi is optional and often partial. Examining the phonetic facts of another famous post-lexical rule, palatalization sandhi in English (e.g., mɪs ju → mɪʃu), Zsiga (1995) finds that a phonetically-based explanation (overlapping gestures in the style of Articulatory Phonology) is more empirically adequate than one involving a phonological rule that changes *s* to *ʃ*. Similar results can be found in the area of vowel sandhi. Traditional phonological accounts of Igbo give identical treatment to intra-word [ATR] harmony (causing, e.g., the alternation of the imperative suffix in *ṣi-ạ* 'tell!' vs. *si-e* 'cook!') and inter-word vowel assimilation (e.g., *nwoke a* → *nwoka a* 'this man'). Both are assumed to be results of feature spreading, differing only in whether they apply at the lexical or postlexical level. But, as Zsiga (1997) shows, while intra-word harmony does have the kind of result we would expect from the application of a categorical phonological rule, the phonetic results of inter-word assimilation continue to show the effects of the apparently "deleted" vowel, in a manner which is again better explain by a gestural account than by a phonological rule.

In his response to Zsiga (1995), Scobbie (1995) poses the question: "What do we do when phonology is powerful enough to imitate phonetics?" It is worthwhile to explore more restrictive models, where phonology does not have the power to imitate phonetics. One of the simplest models of the phonology/phonetics interface has an absolute division between categorical and gradient: all categorical changes are made by phonological rules in the lexical component, while all gradient changes are the result of well-defined phonetic processes (e.g., gestural overlap) that affect the articulatory implementation of the phonological surface representation. While this hypothesis may eventually prove too strong, it is useful to entertain it for the time being, if only as the motivation for a search for counterexamples. The most convincing kind of counterexample would be a case of a categorical postlexical rule — a rule which shows all the signs of the absolute change expected of phonological rules but which could not possibly have applied during the lexical component. If Cree external sandhi really is as it has been described, it would be such a counterexample.

In this study, I look at whether the external sandhi phenomena of Cree are categorical, and thus inconsistent with the simple division of labour described above or whether it is gradient, like vowel assimilation in Igbo, and would be better handled by a theory of phonetics. In short, is the first vowel really deleted?

2. Method

The text analysed is an extract from an autobiographical narrative (Minde 1997) told by Emma Minde, a native speaker of Plains Cree. The linguist present and recording the narrative was Freda Ahenakew, also a native speaker of Plains Cree. The text was transcribed by Ahenakew and H.C. Wolfart, whose decisions as to which vowels have undergone apparent sandhi are followed here.

The analysis focuses on two complementizer particles used at the beginning of the verbal complex in subordinate clauses:

(8) *ê* used in, e.g., temporal subordinate clauses
kâ used in, e.g., relative clauses

The complementizer *ê* can participate in sandhi when it occurs after a vowel-final word or before a vowel-initial adverbial particle or verb stem. We will focus on a single type of case here, that where the complementizer particle merges with the final short *a* of a preceding word, apparently resulting in the deletion of the *a*.

(9) mîna + ê
êkwa mîn ê - miyo - sîhkimât
and also COMP - well - he:encouraged
'and he also encouraged them in the right way'

The complementizer *kâ* can participate in sandhi when the following adverbial particle or verb stem begins with a vowel. The cases we will focus on here involve *kâ* plus a following short *i*, which appear to coalesce into a long *ê*.

(10) kâ + isi
tânisi k - êsi - pimâcihocik
how COMP - thus - they:make:a:living

The three contexts which were measured are summarized in (11).

(11) Class 1: *ê* in a sandhi context after underlying /a/
a) transcribed as having undergone sandhi [...C ê]
b) transcribed with two separate vowels [...Ca ê]
Class 2: [kê] resulting from /kâ/ + following short /i/
Class 3: [ê] in a non-sandhi context, /C#_C/ -- control

Class 1 consists of those contexts where there is a possibility of sandhi between a final short *a* and the complementizer *ê*, as illustrated in (9). Class 1 is subdivided according to whether the transcribers felt that the potential sandhi had or had not occurred. Class 2 consists of those contexts where the complementizer *kâ* merges with a following short *i*, as illustrated in (10). Class 3 consists of control vowels, i.e., instances of the complementizer *ê* which have consonants on both sides and could not possibly have undergone sandhi.

Tokens of each of these three vowels were identified in the text. A token was used only if it did not stand at the beginning of an intonational phrase or within two syllables of the end of an intonational phrase. Tokens were only included in Class 1 if the complementizer *ê* was followed by a consonant, not another vowel.[2]

The intonational phrases containing each token were digitized at 10 000 Hz and the vowel tokens analyzed using a Kay CSL. LPC formant histories were generated for the duration of each vowel token. The formant tracks were edited manually to remove spurious formant readings and give a continuous F1 track. (In cases where the LPC algorithm gave two plausible values for F1, the value with the lower bandwidth was chosen.)

The measurements made for each vowel token were:

---

[2] A possible segmental confound should be mentioned. Most Class 1 tokens occurred in the consonantal context of n_k — the most common *a*-final words undergoing sandhi being *mîna* 'also', *mâna* 'used to', and *anima* 'that one'. Most Class 3 tokens occur in the context of C#_k, where the preceding C is an obstruent.

(12) a) length of vowel
b) frequency of F1 at start of vowel
c) slope of F1

Frequency of F1 at the start of the vowel as determined by the average of the second ten frames of the formant track (that is, the LPC-estimated F1 value at each millisecond between 11 msec and 20 msec after the onset of the vowel). The overall slope of F1 was the slope of the least-mean-squares regression line fitted to the F1 track.

If external sandhi is a true phonological rule, we should expect to find the following:

(13) Predictions if sandhi is phonological
a) No length differences between Class 1a and 3
b) No differences in initial F1 between Classes 1a, 2, and 3
c) No differences in F1 slope between Classes 1a, 2, and 3

If the merger of *a* + *ê* into *ê* in Class 1 sandhi is the result of a true phonological rule that deletes the final short *a*, then the surface representation resulting from this rule will be exactly the same as one which never had an *a* in the first place (e.g., the control tokens of Class 3). The *ê* tokens of Classes 1a and 3 should be indistinguishable in terms of length, intial F1 frequency, and overall F1 slope. We have similar (though weaker) expectations for sandhi Class 2 compared to control Class 3: if a phonological rule has truly coalesced *â+i* into *ê*, then this resulting *ê* should be comparable to the control tokens of Class 3, although the differences in consonantal and syllabic context might result in slight differences.

On the other hand, if sandhi is not a phonological process, we might expect to find the following:

(14) Predictions if sandhi is not phonological
a) a longer vowel in Class 1a & 2 than in Class 3
b) a higher initial F1 in Class 1a & 2 than in Class 3
c) a (more) negative slope for F1 in Class 1a & 2

For Class 1, if no phonological rule has deleted the *a* of an *a+ê* sequence, then we would expect to see the effects of this *a* (however weakened) in the phonetic realization of the token. We would expect an *a+ê* sequence to be longer than a simple *ê* and to have a higher initial F1 value and an overall negative F1 slope as the token moves from something closer to an [a] to its final [e:]. We might also expect there to be no clear-cut distinction between those a+ee sequences which Ahenakew and Wolfart transcribed as having undergone sandhi (Class 1a) and those transcribed as not having undergone sandhi (Class 1b).

3. Results

The results of the measurements are summarized in Table 1. Boxplots for the major results are given in Figures 1-3.

	Class 1a sandhi [ê]s	Class 1b non-sandhi [a ê]s	Class 2 sandhi [kê]s	Class 3 control [ê]s
N	44	16	42	41
Length mean	0.0837	0.1008	0.0732	0.0679
s.d.	0.0499	0.0516	0.0336	0.0199
Starting F1	549.5	588.5	462.7	489.4
s.d.	67.47	83.23	62.7	58.5
Slope mean	-0.5817	-0.7671	0.0327	-0.0714
s.d.	0.4393	0.6855	0.3369	0.4821

Table 1. Results

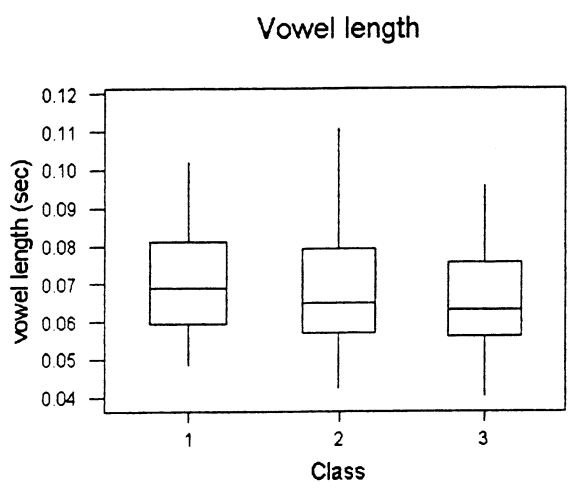


Figure 1

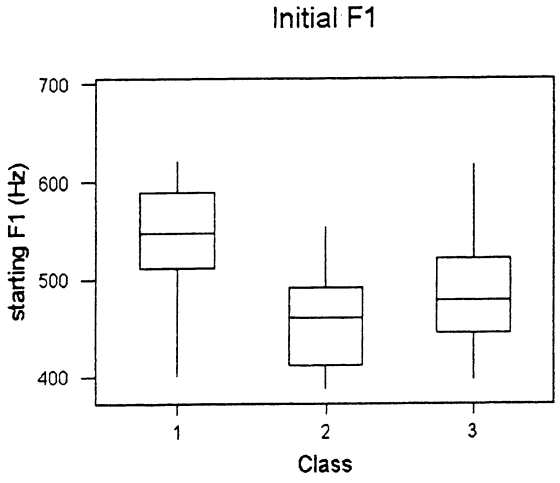


Figure 2

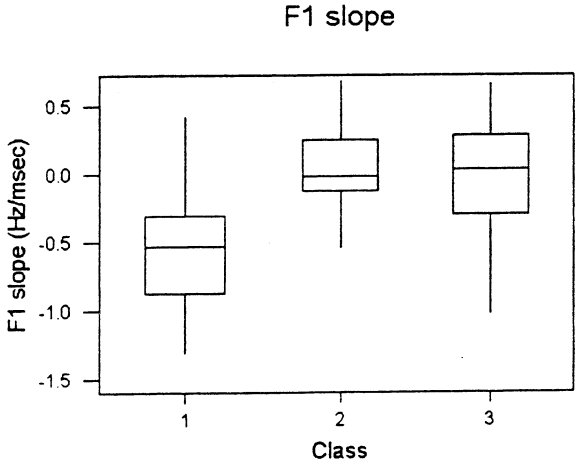


Figure 3

Comparing Class 1a against Class 3, those [e:]s which have undergone sandhi are longer than the control [e:]s (84 msec on average compared to 68 msec for the controls), but the difference is not quite significant ($F = 3.59$, $p = 0.062$). The frequency of F1 at the beginning of the vowel is significantly higher in sandhi [e:]s than in control [e:]s ($F = 19.12$, $p < 0.001$). The overall slope of F1 was significantly more negative in the sandhi [e:]s than in the control [e:]s ($F = 26.08$, $p < 0.001$).

Comparing Class 2 (kâ+i→kê) against the control class, there was no significant difference between the two in terms of either length ($F = 0.76$, $P = 0.387$) or F1 slope ($F = 1.31$, $p = 0.257$). There was a barely significant difference between the two classes in the initial frequency of F1 ($F = 3.99$, $p = 0.049$). This difference was, however, the *opposite* of the predicted difference — the initial F1 of *kê*'s resulting from *kâ+i* is *lower* than that of the control class, that is, the start of a [ke:] < /ka:+i/ is less like [a] than is an underlying [e].

4. Conclusion

Let us summarize by comparing the results just described with the expectations outlined in (13) and (14).

(17) If the underlying /a/ or /â/ has not really been deleted, we should find:
- a) a longer vowel in Class 1a (a#e) and 2 (a+i>e) than in Class 3 (control)
  - Class 1a tokens are longer than Class 3, but not quite significantly so.
  - Class 2 is not longer than Class 3.
- b) a higher initial F1 in Class 1a and 2 than in Class 3
  - Class 1a has a significantly higher initial F1 than Class 3.
  - Class 2 has a barely significantly *lower* initial F1 than Class 3.
- c) a (more) negative slope for F1 in Class 1a and 2
  - Class 1a has a significantly greater negative slope than Class 3.
  - Class 2 does not.

The conclusion we are forced to is that the final *a* in Class 1a has not really been deleted. If it had been, the resulting *ê* should have been indistinguishable from the underlying *ê* of the control class. Certainly, the word-final vowels are significantly changed from what they would be in a non-sandhi context: compare the initial F1 of 550 Hz for final [a]s in the sandhi context of Class 1a with the typical inital F1 of 686 Hz for unstressed phrase-medial word-final *a* in non-sandhi contexts. But this change cannot be attributed to a phonological deletion of the word-final /a/ segments. The articulatory target for an [a] is clearly still present and, while the articulators might seldom reach that target in a sandhi context during fluent speech, the target continues to influence the overall course of the articulators.

Things are different for Class 2, those tokens of *ê* arising from complementizer *kâ* plus a following *i*. The results suggest that these tokens truly have undergone a phonological rule which has resulted in a single [e:] target. These tokens are indistinguishable from underlying /e:/s in all respects, except for initial F1, where the [e:] resulting from /a:+i/ coalescence is even *less* like [a] than underlying /e:/s are.

This unexpected significant difference in initial F1 between Class 2 and Class 3 might be attributable to the segmental confound mentioned in footnote 2. Another possible explanation is the role of phonological contrast. Wolfart (1989) notes that the coalescence of *kâ+i* into *kê* does not seem to be truly optional. Rather, it is nearly exceptionless when the verb occurs in a realis clause, and is not found in irrealis clauses. If so, the contrast between

[ka:i] and [ke:] is semantically significant, and it is not implausible that speakers might exaggerate the difference, or rather, that speakers will take more care to approach the [e:] target in those cases where it is potentially contrastive (Class 2) than in those cases where there is no danger of confusion (Class 3).

We have a clear difference between Class 1a and Class 2 kinds of sandhi in Cree. If our ultimate goal is to do away with the need for phonological juncture rules, then Class 1a is no problem. It turns out that there is no need for a phonological rule to turn the sequence *mîna ê* into *mîn ê*. Indeed, an analysis which did make use of such a phonological rule would be empirically inferior.

But Class 2 sandhi cannot be dismissed so easily as a mere apparent quirk of phonetic interpretation. It is noteworthy, however, that Class 2 sandhi is very unlike the kind of across-the-board rules that would pose the greatest challenge for a simpler phonological theory. It was odd in the first place for being the only example of sandhi where the first vowel was long rather than short. Its application is restricted a single complementizer particle *kâ*. Its context of application is also restricted: the following short *i* will always occur in a stem or a preverb that is built on the root *it-/is-* 'thus, so'. Verbs built on this root in Cree for a class whose members share other interesting morphosyntactic properties. And, even when coalescence does occur between this single complementizer particle and this restricted class of verbs, its occurrence seems to be semantically governed and correlates with a realis modality. In short, while this coalescence is categorical, it is more comparable to the contraction of English *will not* to *won't* than to an across-the-board structure-changing sandhi phenomenon that would require a massive complication of the phonology-phonetics interface.

References

Bloomfield, Leonard (1930); *Sacred Stories of the Sweet-Grass Cree*; National Museum of Canada, Bulletin 60.

Bloomfield, Leonard (1934); Plains Cree Texts; Publications of the American Ethnological Society, 16.

Condoravdi, Cleo (1990); "Sandhi rules of Greek and prosodic theory"; in Sharon Inkelas and Draga Zec, eds., *The Phonology-Phonetics Interface*; Chicago: The University of Chicago Press.

Hayes, Bruce (1990); "Precompiled phrasal phonology"; in Sharon Inkelas and Draga Zec, eds., *The Phonology-Syntax Connection*; Chicago: The University of Chicago Press.

Kaisse, Ellen (1985); *Connected Speech: The Interaction of Syntax and Phonology*; San Diego: Academic Press.

Minde, Emma (1997); *kwayask ê-kî-pê-kiskinowâpahtihicik: Their Example Showed Me the Way*; told by Emma Minde; edited and translated by Freda Ahenakew and H.C. Wolfart; Edmonton: The University of Alberta Press.

Scobbie, James (1995); "What do we do when phonology is powerful enough to imitate phonetics: comments on Zsiga"; in Bruce Connell and Amalia Arvaniti, eds., *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*; Cambridge University Press.

Wolfart, H.C. (1973); *Plains Cree: A Grammatical Study*; American Philosophical Society transactions.

Wolfart, H.C. (1989); "Prosodische Grenzsignale im Plains Cree"; *Folia Linguistica*, 23: 327-334.

Zsiga, Elizabeth C. (1995); "An acoustic and electropalatographic study of lexical and postlexical palatalization in American English"; in Bruce Connell and Amalia Arvaniti, eds., *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*; Cambridge University Press.

Zsiga, Elizabeth C. (1997); "Features, gestures, and Igbo vowels: An approach to the phonology-phonetics interface"; *Language* 73, 2, 227-274.

The phonetic word: the articulation of stress and boundaries in Italian

Edda Farnetani
Istituto di Fonetica del CNR
Padova (Italy)

1.Introduction

A spoken word produced in its canonical form can be defined as a phonetic entity with its specific segmental and suprasegmental structure. The suprasegmental component is expressed by two sets of acoustic/perceptual attributes: temporal/melodic properties at the word's edges, signaling initial and final boundaries, and, for polysyllabic words, prominence contrast between syllables. In free-stress languages prominence at the word level is determined by the position of lexical stress.
In normal speech communication, the word, unless uttered in isolation or produced in clear speech style, tends to lose its canonical properties and to undergo a number of changes both in the segmental and the suprasegmental properties. Such changes usually referred to as connected speech processes (see Kohler, 1990 for a review), range from segmental reductions, weakening of stress contrast, assimilations, to feature and segment deletion, cancellation of boundary signals and of lexical prominence. In this case the word disappears as a phonetic unit. A great number of studies show that speaking rate and speech style are two important factors affecting the articulation of a segment (see Perkell, 1977, p.352 for a review of the changes in the articulatory kinematics as a function of these two factors); linguistic factors such as prosodic/syntactic structure have traditionally been investigated in terms of acoustic melodic and/or durational effects. Only recent research has been directed to the analysis and the modeling of the changes in the articulatory configuration and in the kinematics of gestures as a function of prosodic structure (see below, 2.1 for a review of recent studies on English and French).
This experiment, which expands previous research on this topic (Farnetani and Vayra, 1996) deals with the supraglottal articulation of lexical stress and of word boundaries, and tests the effects of three different speech contexts: isolated words, sentence final words, and words embedded in a three-component syntactic phrase. Thus it compares citation forms with words in two syntactic/prosodic positions, phrase/sentence final position, and phrase-medial position. The 1996 study dealt with the spatial effects of stress and boundaries, the present paper is concerned with both the spatial and the temporal/dynamic aspects of stress and boundary production.
One purpose if this study is to assess how the articulation of vowels and consonants in CV syllables signals prominence and initial and final boundaries, and whether and how the articulation changes as a function the three prosodic contexts; a second purpose is to get more knowledge of the dynamics of stress and boundary production, in the light of current phonetic models of their realization.

1.1. Previous studies on lexical stress and boundaries in Italian.

As for boundaries, the fact itself that in continuous speech words combine together to form larger prosodic units implies that at the lexical level boundary signals are weaker than at any higher level. The question is then: are word level boundaries preserved in items embedded in larger prosodic units, and if so, to what extent are they weakened? Or are they totally canceled?. As for lexical stress, we can ask whether the syllabic prominence associated with stress weakens in embedded words. A previous acoustic study on boundaries in Italian (F0 and duration) indicated that within two-word phrases, both final boundaries of W1 and initial boundaries of W2 disappear, while syllabic prominence is preserved in both words (Farnetani, 1989). Other acoustic studies on lexical stress and rhythmic structure (Farnetani and Kori, 1983, 1990) indicate that in phrase-non-final bisyllabic words lexical prominence is preserved, but that stressed vowels undergo a progressive durational shortening as the interstress interval between the target word and the following one decreases. The 1983 study shows that word level prominence is totally canceled only under the condition of stress clash.
The overall data clearly indicate that in continuous speech word level prominence resists reduction and deletion more than boundary signals, and therefore becomes itself a cue to word parsing. This idea is

supported by a number of ASR studies (Lea, 1980), where the function of lexical prominence for detecting lexical items in continuous speech has been tested for English: the results show that prominent syllables are not only anchor points for phonetic recognition, but also a guide to the listeners for word detection, even if they do not, *per se,* delimit words from the neighbours.

1.2. Current models of stress and boundaries production

Three models of the production of prominence have recently been proposed: the "Jaw expansion" model of Macchi (1985), the "Sonority" model of Beckman, Edwards and Flechter (1992), the "Hyperarticulation" model, based on Lindblom's Hyper/Hypospeech theory (1990), and proposed by de Jong (1995) to account for the articulation of prominence. Although quite different, all three of them seem to account for the main acoustic effects of prominence observed in English and in other languages: longer durations, more extreme vowel-formant frequencies, and higher intensity.

The jaw expansion model attributes the acoustic effects of stress to an expansion in space and time of jaw movement cycles, i.e. a greater and longer opening for vowels, and a more extreme and longer closing for consonants.

According to the sonority model, framed within the task-dynamic model of speech production (Browman and Goldstein (1989), it is the whole vocal tract to expand its cycles in space and time, thus both jaw, tongue, and lip movements increase in magnitude and duration in the production of prominence. Expansion is brought about by decreasing the temporal overlap between adjacent gestures, and aims at increasing the contrast along the scale of the sonority feature. Thus, making a syllable prominent means to maximize the sonority contrast between C and V of the syllable.

The H&H theory, based on the movement-target model of speech production, assumes that speech varies along a dimension from hypo- to hyper-articulation: at one end, hyperarticulation i.e. target overshoot aims at maximizing the segment phonetic distinctiveness, and is proper of clear speech style, at the other extreme, hypoarticulation i.e. target undershoot, aims at maximizing motor economy and characterizes colloquial, casual speech.

de Jong (1995) tested the three models, by analyzing the movements of the jaw, the tongue and the lips in monosyllabic words with nuclear accent (specifically in narrow focus condition), vs words with prenuclear accent, vs unaccented words. The observed tongue and (lower) lip positions were decomposed into independent jaw, tongue and lip components, which made it possible to infer the active movements of the tongue and lips. The results show that for nuclear /U/ the tongue articulator is indeed active in the production of prominence (contrary to the jaw expansion model), but its movements do not only expand along the high-low dimension, but also along the front-back dimension, suggesting a more posterior target specification; and the upper lips enhance their protrusion, as compared to the prenuclear and the unaccented /U/. Moreover the kinematics of gestures shows that the movement expansion in nuclear words is associated more strongly with an increase in velocity than in duration. According to de Jong, the hyper/hypospeech model is the only one that can account for all the findings, therefore nuclear prominence has to be considered an instance of localized hyperarticulation.

It must be noted that study by de Jong compares emphatic with prefocal accent with no accent, while the material used by Edwards, Beckman and Fletcher (1991); Beckman et al. (1992) compares phrase-final accented words with phrase nonfinal unaccented words, hence, the levels of prominence specification compared in the two studies are not the same.

As for final prosodic boundaries, the experiment by Edwards et al.(1991) proposes two alternative production strategies for the articulation of final boundaries: slowing down the speech tempo (by changing gestural stiffness), decreasing gestural overlap (by changing intergestural phasing). Both strategies bring about final syllable lengthening, but the second has also the effect of expanding the articulatory gestures and is applied in most cases for unstressed final syllables.

For the articulation of initial boundaries the EPG data for English by Keating (1995; in press) indicate that the tongue-to-palate contact in alveolars is always greater in word-initial than in word-final position and that in sentence initial words the contact of initial Cs is greater than in words in other positions. In line with de Jong's hyperarticulation hypothesis, Keating proposes that consonant strengthening in initial boundaries is an instance of hyperarticulation. The EPG data by Fougeron and Keating (1996) for French refer to tongue-to-palate contact for /n/ and /t/ production, and show that, at the word level, initial Cs have more contact than word medial Cs, and that the contact in initial consonants increases gradually as the prosodic level becomes higher, the greatest contact being associated with utterance initial position.

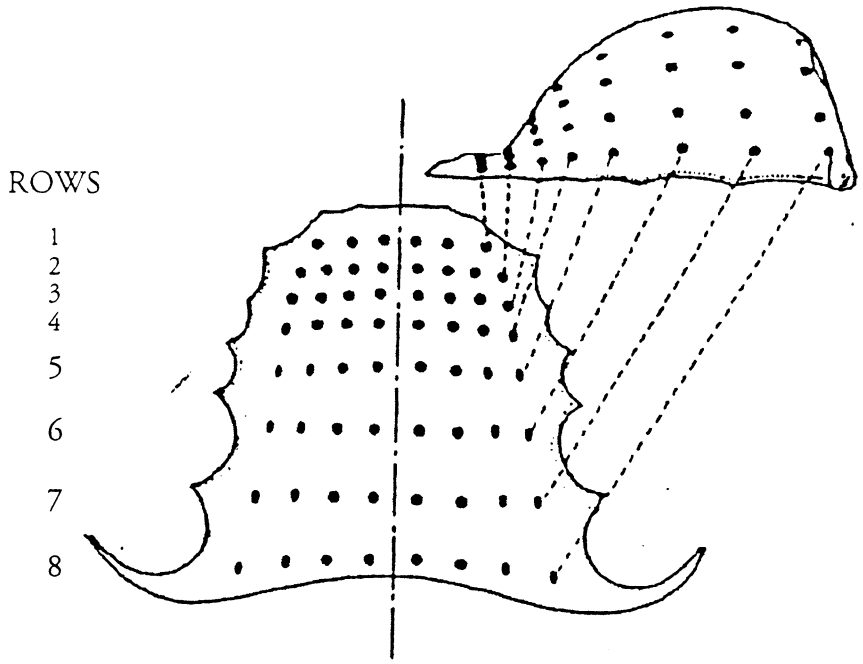


Fig. 1: Artificial palate: the electrodes are distribuited in eight rows from front to back.

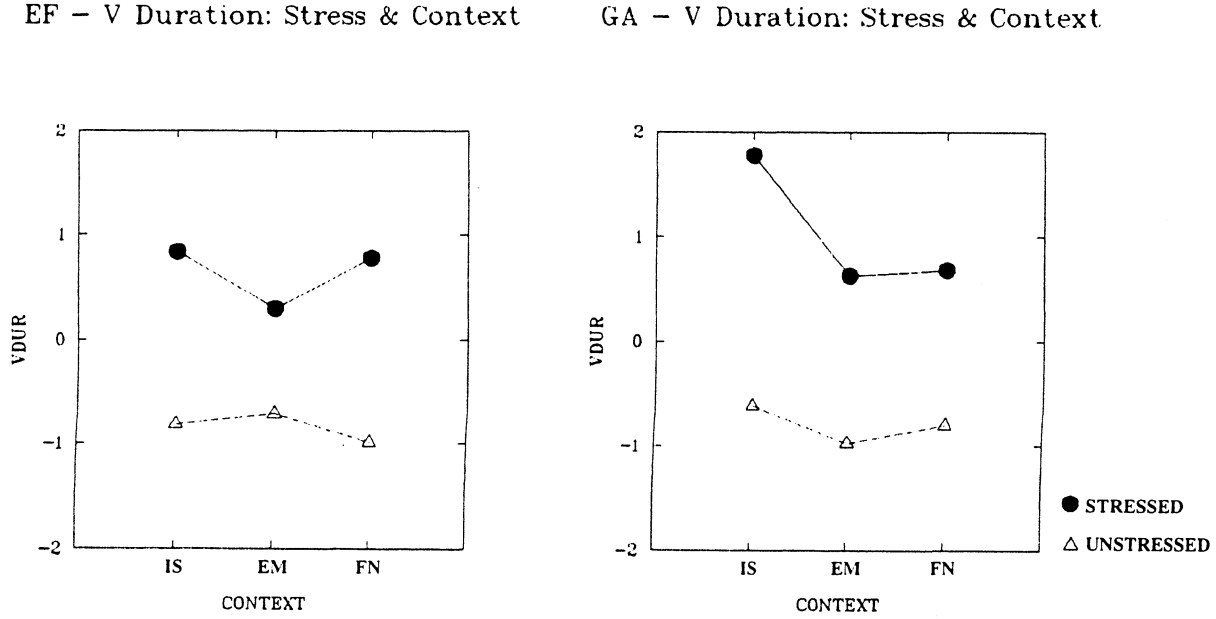


Fig. 2: Vowel durations, averaged across /i/, /a/, /u/, as a function of stress and context

2. Methods

2.1 Speech Material

The patterns of tongue movements and configurations in the production of stress and boundaries were investigated by means of EPG in trisyllabic CVCVCV words with varying lexical stress position. where C=/t/ and V= /a/ or /i/ or /u/ in symmetric sequences. The target words are the reiterant versions of the proper names *Giacomo, Nicola, Niccolò* (stress on syll.1, 2, and 3 respectively). The words were produced in isolation (IS), in sentence final position (FN) (*Partì per la Francia col marchese Ugo...*), and embedded in a three-constituent phrase in sentence initial position (EM) (*Ugo...della Torre partì per la Francia).* The utterances were produced with no emphatic stress on any of the words: the isolated words and the sentence final words were produced with the expected sentence accent, and the embedded words with no accent or a weak accent (which preceded the phrasal accent of the phrase final word).

The material was repeated 5 to 6 times at the most comfortable reading rate by two Northern speakers of Standard Italian (EF and GA).

The use of trisyllabic words has allowed independent analysis of stress and boundaries: the effects of stress were analyzed paradigmatically in word-medial syllables, those of boundaries were analyzed by comparing word-medial syllables with word-initial and word-final ones (with both paradigmatic and syntagmatic comparisons).

The data were collected by means of the Reading EPG system which simultaneously records EPG and acoustic data at 10 ms rate. See, in Fig. 1 the distribution of electrodes along eight rows ordered from front to back.

2.2. Measurements and analyses

For the consonant the data were taken in the EPG frame of maximum front contact in the area extending from row 1 to row 5, which covers the alveolar and the prepalatal region, and the parameter is CFRONT, i.e. the percent electrode activation in this region. This large area was chosen in order to avoid ceiling effects in the articulation of consonants in prominent syllables.

For the vowels the measures were taken in the frame of maximum contact for /i/ and in that of minimum contact for /u/ and /a/. The vowel parameters are: PI (posteriority index), CI (centrality index).

PI indicates the variations in contact along the front/back dimension and is expressed by the row number where the frontmost contact occurs: PI ranges from 1, when the first row is contacted, to 9, when no contact occurs in any of the rows (which sometimes occurs for stressed /a/). Normally, the frontmost contact in the production of vowels is a contact at the sides of the palate. The right/left asymmetries were accounted for by adding to the row number a fixed value of 0.50 when the contact was on one side only.

CI, taken in the palatal and postpalatal region (Row 6 to 8) is a measure of the extent of tongue contact from the sides to the center of the palate, and is expressed by the number of activated electrodes in the row of maximum contact. CI ranges from 0 (no contact) to 8 (when all the eight electrodes of the row are contacted), and is corrected by subtracting a fixed value of 0.33 or 0.66, when one or two of the back rows, respectively, exhibit less central contact than the one with maximun contact. As can be seen in the palate profile of Fig.1, in this area an increase in centrality indicates a movement of the tongue body to a higher position.

Temporal parameters are: C closure duration, acoustic vowel duration.

In order to shed light on how the various changes are brought about in the CV syllable, the following measures were taken for /ta/ syllables: 1) MaxC (n. activated electrodes for /t/ in the last frame of max closure; 2) Vmin (n. of free electrodes in the first frame of minimum contact for /a/); 3) duration of the interval between CMax and Vmin; 4) difference in contact, and 5) mean and max rate of contact change between CMax and Vmin. The last three measures should reflect the amount of tongue displacement, its duration and rate in the C-to-V gestures The present paper will refer only to the very preliminary results of this analysis.

Comparisons among parameters were carried out on standardized values. The significance level of statistical analysis (ANOVA) was set at $p<0.02$.

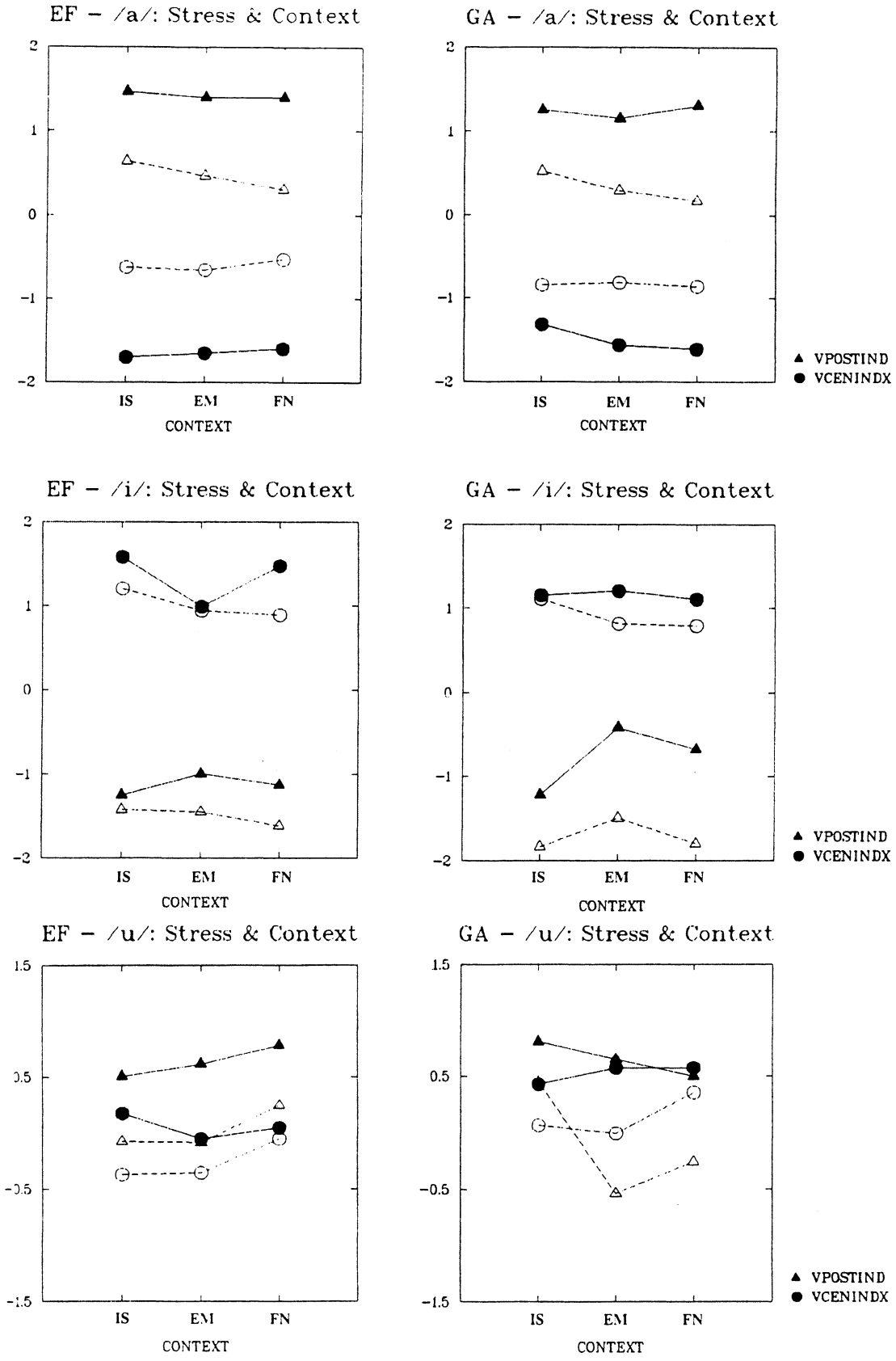


Fig. 3: Mean values of PI and CI for /a/, /i/, and /u/, as a function of stress and context

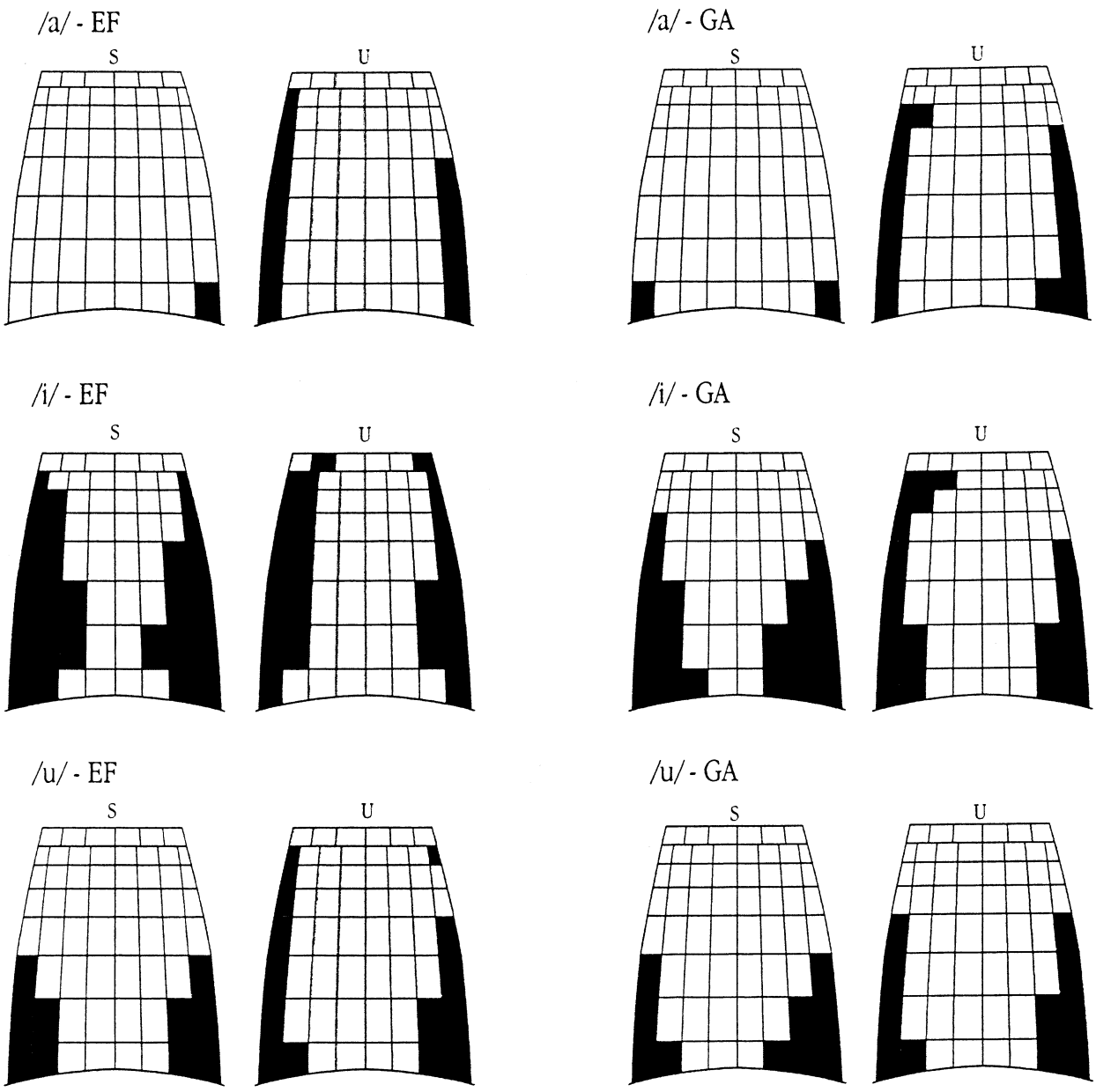


Fig. 4: Typical palatographic configurations of stressed and unstressed vowels.

3. Results

3.1. Prominence

3.1.1. Prominence in vowels

The results of comparisons between word-medial stressed and unstressed syllables, across the different prosodic conditions indicate that stress has significant effects on both the duration and the spatial configuration of all three vowels. The direction of the changes is the same for the two subjects. The data are illustrated in Fig.2, which shows the acoustic durations (averaged across the vowels), and Fig.3, which shows the spatial parameters for each vowel, in each prosodic context.

As for duration, stressed vowels are always significantly longer than unstressed ones, as expected (see Fig.2). The average durations are: ms 113.67 (S), ms 56.05 (U). Also the prosodic context has systematic effects on vowel duration for both subjects. Interactions between context and stress indicate that only stressed vowels change significantly as a function of context. As illustrated in Fig.2, for EF the stressed Vs are shorter in embedded words than in the other two conditions (ms 126.71 (IS), ms 105,20 (EM), ms 124.36 (FN), while for GA, the stressed Vs of embedded and final words are both shorter than the stressed Vs of isolated words (ms 130.60 (IS), ms 96.00 (EM), ms 97.76 (FN).

As for the spatial configurations (see Fig.3), stressed /a/ is produced with a more posterior, a more peripheral, and globally less tongue body contact than unstressed /a/ (+PI and -CI) for both subjects. Stressed /i/ and /u/, like stressed /a/, are produced with a more posterior tongue contact (+PI). The averaged PI values are: PI= 6.04 (S), PI=4.32 (U). In high vowels, however, also the centrality index tends to increase with stress (+CI). The trend to increase CI in stressed high vowels is consistent, although it does not reach the significance level in all three prosodic contexts (see for EF, the data for /i/ in embedded words (EM), and for /u/ in final words (FN) in Fig.3). Typical palatographic configurations of stressed and unstressed vowels for the two subjects are illustrated in Fig.4.

The context has no systematic effects on the vowel spatial parameters. Specifically, there is no evidence of a consistent trend towards a weakening of the articulatory characteristics of prominence in the stressed vowels of the weakly accented embedded words vs isolated or final words (see, Fig. 3, vowel /a/ for both subjects). There are some effects for the high vowels in the expected direction (reduction of articulatory prominence in embedded words) but they are not shared by the two subjects. For GA posteriority of both stressed and unstressed /u/ decreases in embedded words ($p=0.015$). For EF, the only effects of context ($p = 0.021$) occur in stressed /i/, where centrality decreases from CI= 4.7 (isolated words) to CI= 3.9 (embedded words).

3.1.2. Prominence in Consonants

The data on the duration and the maximum EPG contact for the key consonant /t/ are shown for each subject in Fig.5. The analysis indicates that both stress and prosodic context affect the EPG contact and the duration of the consonant. For both subjects CFRONT and C duration are larger in stressed than in unstressed syllables and larger in isolated and final words than in embedded words (p values ranging from 0.001 to 0.000). The average effects of stress are, for CFRONT: 65.9% (S), 57.5% (U), for CDUR: ms 84.63 (S), ms 54.67 (U). The average effects of context are, for CFRONT: 71.5% (IS), 60.97% (FN), 52.70% (EM); for CDUR: ms 91.37 (IS), ms 68.07 (FN), ms 49.51 (EM). Fig 5 clearly shows that stress and context sum up their effects so that consonants in stressed syllables in isolated words have the maximum contact and duration, those in unstressed syllables in embedded words have the minimum. In spite of the wide range of front contact variation, cases of reduced and heavily reduced consonants (from incomplete closure to side contact) were observed only in unstressed syllables of embedded and final words; they amount to 8% for EF and 13% for GA.

3.1.3. Comments on the articulation of stress

The systematic increase in PI (i.e. decrease in anterior side contact) in all three stressed vowels as compared to the unstressed ones, indicates that they are produced with a lower tongue position in front of the constriction. While in stressed /a/ CI decreases (indicating that also the tongue body is lower in stressed than in unstressed /a/), in the production of stressed /i/ and /u/ CI increases, indicating a higher tongue body in the palatal and postpalatal regions. Thus, increase in both parameters (+CI, +PI) indicates that stressed high vowels are articulated with a lower tongue front, and a higher tongue body than the unstressed counterparts. Even if it would be possible to interpret the two sets of data as movements of the tongue proper, i.e. retraction of tongue front associated with elevation of tongue body, the fact that in stressed /a/ the increase in PI is not accompanied by any tongue body elevation, and that in high vowels the increase in PI is larger and more systematic than the increase in CI, is

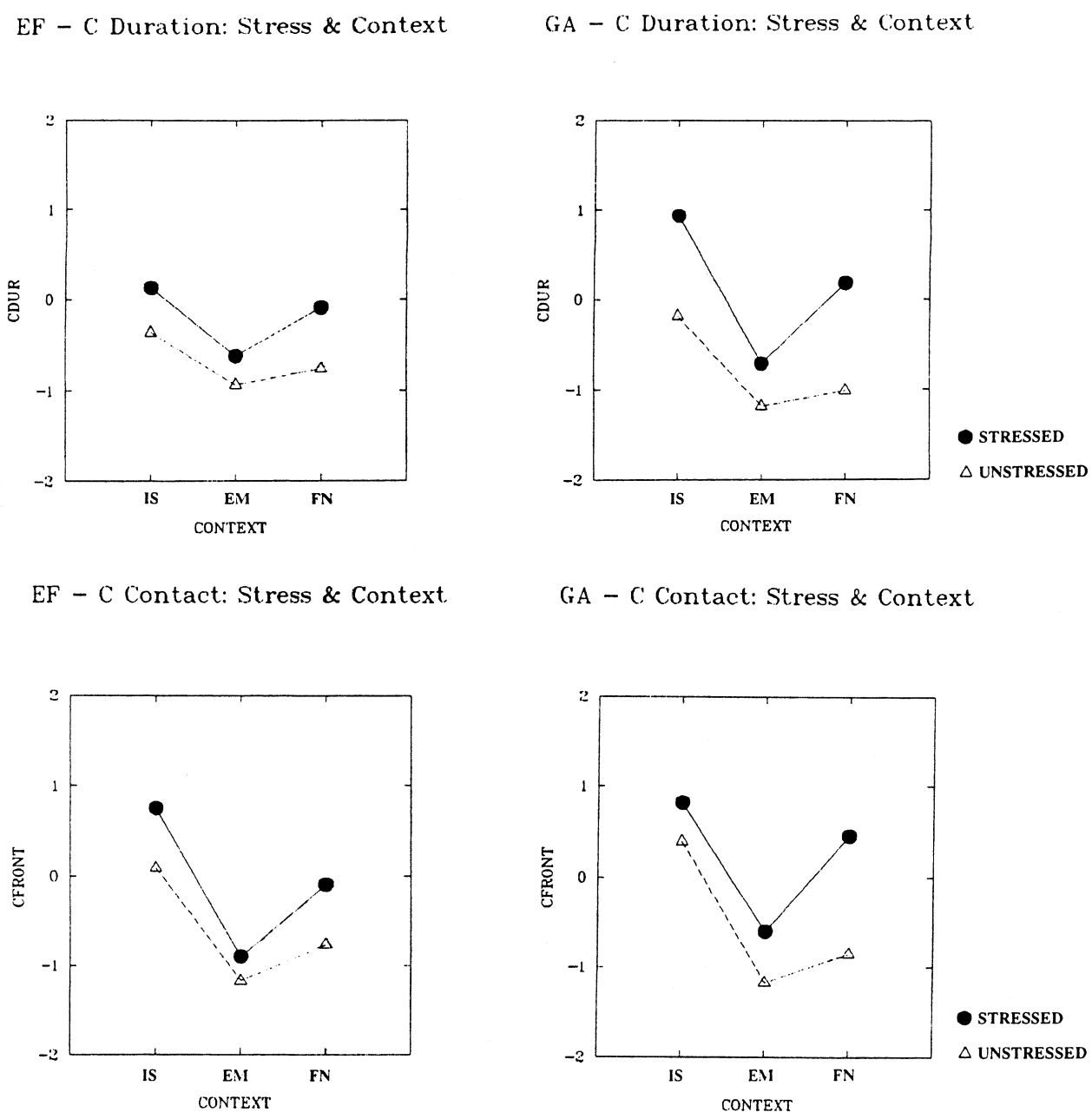


Fig. 5: Mean durations (upper graphs) and mean values of Front Contact (lower graphs) of consonant /t/ as a function of stress and context.

against this hypothesis. Studies on the role of the jaw and of the tongue in the production of prominence (Edwards et al. 1991, de Jong, 1995, among others), and a pilot Movetrack study on vowels /i/ and /a/ in Italian for subject EF (Farnetani and Faber, 1992) lead to an account of the present EPG findings in terms of tongue and jaw articulation. The 1992 data on Italian showed that both stressed /a/ and /i/ were produced with a lower jaw position than the unstressed counterparts. As for the tongue position, it was lower for stressed /a/, but not for stressed /i/. The separation between the jaw and the tongue component in the tongue data, indicated that the observed tongue position was the result of a larger rising movement of tongue articulator in the production of stressed /i/. The two independent movements of the tongue and the jaw in opposite directions in the production of stressed /i/ converge with the present EPG observation. We can thus infer that the increase in posteriority for stressed /a/, /i/, and /u/ be associated with a jaw lowering movement (much larger for /a/ than for /i/ or /u/), while the increase in centrality around the constriction place of high vowels be associated with a more extended tongue rising movements toward the target, and a consequent tighter constriction.
The effects of the prosodic context on the duration of word-medial stressed vowels, which are shorter in embedded vs isolated words for both subjects, indicate that longer segments are more subject to temporal compression than shorter ones, and that also word medial syllables undergo temporal compression in continuous speech. Moreover, the vowel shortening in embedded words vs sentence final words observed for one subject (EF), indicates either a global lengthening of utterance final words, or according to the model of Lindblom and Rapp (1973), a trend towards a durational compression of words as the distance from the utterance boundary increases. The fact that the stressed vowels in embedded words do not systematically undergo reduction in spite of their shorter durations, suggests the activation of compensatory manoeuvres to preserve the spatial quality characterizing prominent vowels.
The consonants parallel the vowels in the durational and spatial changes as a function of stress: syllable prominence is always associated with a tighter, more expanded, and longer consonant closure. Instead, consonants differ from vowels in the effects of context: stressed consonants reduce both their duration and the tongue to palate contact in embedded words. In this condition stressed consonants may be considered less strong than in isolated words, although it must be reminded that in embedded words they are always articulated with complete closure with a contact (average CFRONT= 56%) covering the three first rows, i.e. the entire dentoalveolar region.

3.2. The articulation of boundaries.

In order to test whether the whole CV syllable or only the word initial and final segments are affected by word and utterance boundaries, both C and V were analyzed. Fig.6 and Fig.7 illustrate the global results for the consonant and the vowels, respectively.

3.2.1. Initial boundaries: consonant /t/

The results of paradigmatic comparisons between word initial and word medial consonants (C1 vs C2) indicate that both the closure duration (CDUR) and the C front contact (CFRONT) are larger in word initial position than in word medial position. The data are highly significant for both subjects, and for each prosodic context (p=.04 0.000), as illustrated in Fig.6 (syll. 1 vs syll.2). Moreover the data indicate that the boundary effects on C are significant in both stressed and unstressed syllables.
The interesting finding concerning initial boundaries is that they are signalled also in embedded and in final words: it is reminded that in both conditions the word initial boundaries are phrase internal, hence it appears that initial boundaries are signalled also at the word level. The average differences between C1 and C2, which can be considered to reflect the strength of initial boundaries in the three prosodic contexts are: CDUR = ms 97.84 (IS); ms 26.21 (EM); ms 23.86 (FN); CFRONT = 8.87% (IS); 14.41 (EM); 10.78% (FN). Thus, it seems that in the more constrained condition of phrase-internal boundaries, it is the strengthening of the closure, more than the increase in duration of the silent interval to mark the word initial boundary (see Fig.6). The reason why the very long duration of consonants in absolutely initial position is not paralleled by a comparable increase in contact can be found in the fact that the contact in isolated words is always very high and cannot extend further back once it has reached the maximum extent compatible with the articulation of an alveolar stop consonant (around 82% of the entire front region).

3.2.2. Initial Boundaries: Vowels

The changes in V1 with respect to V2 suggest that also the vowel in syll.1 may contribute to signal

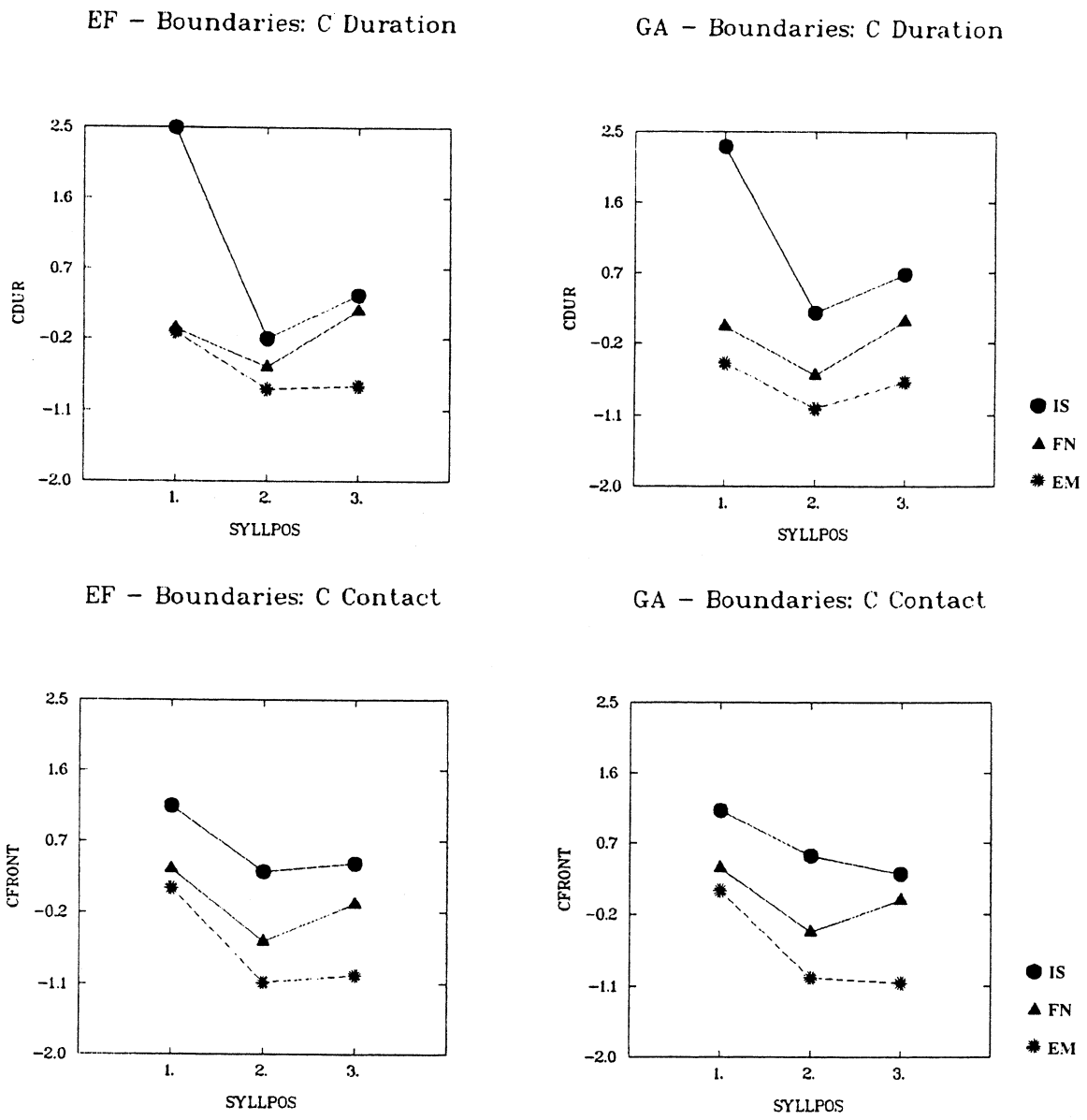


Fig. 6: Mean durations (upper graphs) and mean values of Front Contact (lower graphs) of consonant /t/ as a function of boundaries.

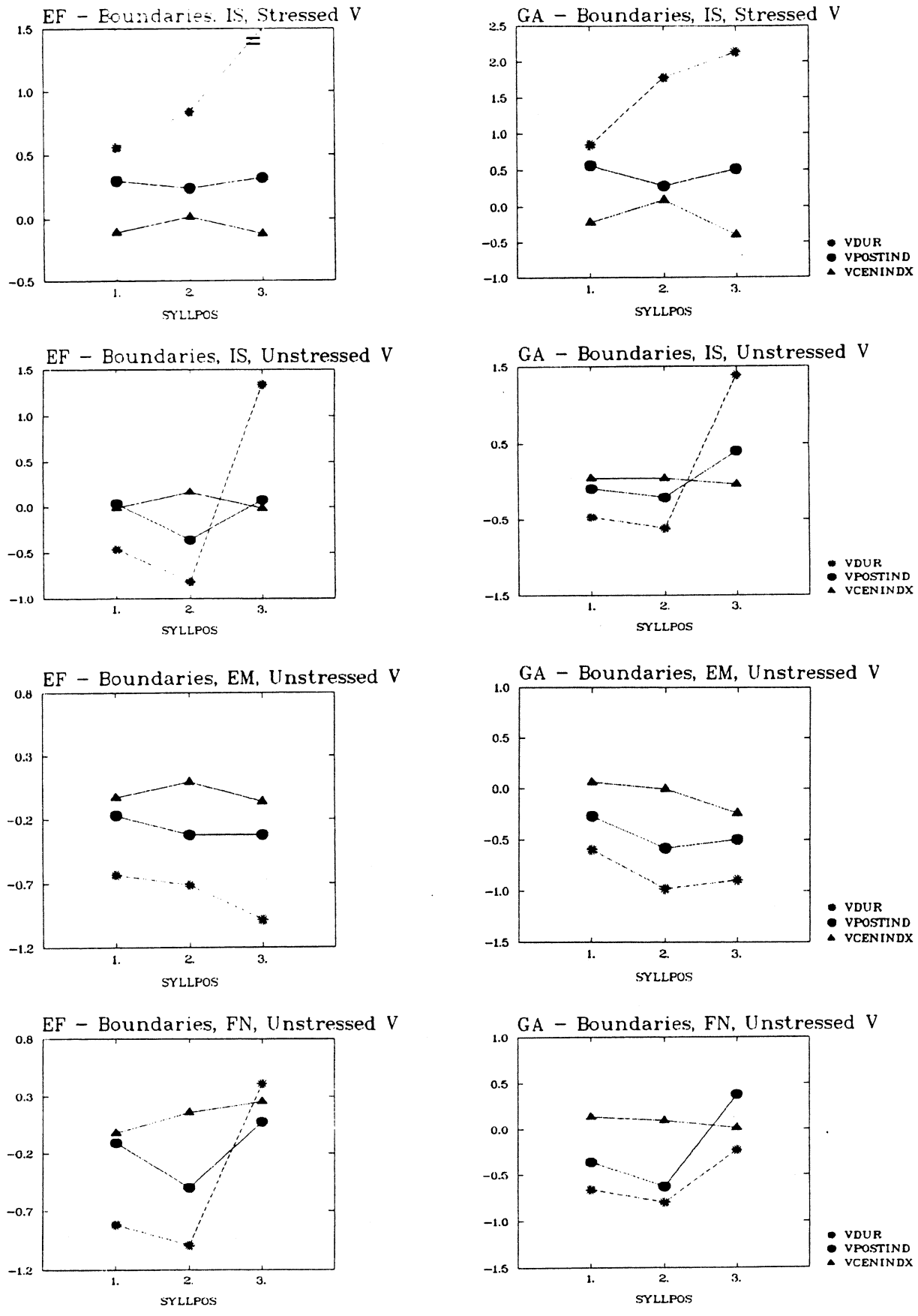


Fig. 7: Average values of vowel duration, PI and CI as a function of boundaries.

initial boundaries. The most systematic spatial change concerns the posteriory index (PI), which is larger in V1 vs V2, and indicates that vowels in syll.1 are produced with an increased opening of the vocal tract.
In isolated words both the stressed and the unstressed vowels tend to have larger PI values syll. 1 than in syll.2, with p=0.000 for EF, p=0.008 for GA (see Fig.7, IS Stressed V; IS Unstressed V, syll.1 vs syll.2, for both subjects). For EF also CI varies significantly and tends to decrease in the first syllable (p=0.013 for stressed Vs; p=.04 0.017 for unstressed Vs). The durational data are peculiar and indicate that in isolated trisyllables stressed vowels in initial position ('CVCVCV words) tend to be shorter than those in medial position (CV'CVCV words) (p=0.010 for EF; p=0.000 for GA), while unstressed vowels are always longer in syll.1 than in syll.2 (p=.04 0.000 for EF; p=.04 0.004 for GA).
In embedded and final words only unstressed vowels vary significantly as a function of boundaries. with PI tending to be higher in V1 than in V2 (p=.04 0.001 for EF; p=.04 0.001 for GA), thus reflecting the trend observed in isolated words. VDUR is higher in V1 vs V2 only occasionally, in embedded words for GA, and in final words for EF (see Fig.7, GA - EM, EF -FN, syll.1 vs syll.2)
The overall data indicate that the vowels of the first syllables (especially unstressed Vs) tend to be more prominent than word medial vowels owing to a larger vocal tract opening often associated with a longer duration. In embedded and final words, these differences (albeit rather small), together with the systematic increase in tongue-front contact of the onset consonant, can contribute to cue to a word-initial boundary in phrase-internal position.

3.2.3. Final boundaries: consonant /t/

The results of comparisons of C2 vs C3 indicate that also the onset consonant of the final syllable contributes to mark final boundaries. Significant changes occur in isolated and sentence final words but not in embedded words. Final boundaries are signalled by significant increase in closure duration in the last syllable : as can be seen in Fig.6, CDUR is longer in syll.3 than in syll.2 (p=0.000 for both subjects). The front contact (CFRONT) varies at a much lesser extent: it tends to be higher in syll.3 than in syll.2 only in sentence final words (p=0.01 for EF; p=0.047 for GA)

3.2.4. Final boundaries: Vowels

The overall results indicate that the word final boundaries are signalled in isolated and in sentence final words, as expected, but not in embedded words. (see Fig.7, Syll.2 vs syll.3 : IS vs EM vs FN for both subjects).
Final boundaries are signalled by a significant increase in duration of final vowels. As illustrated in Fig.7, V3 is longer than V2 for both subjects (p=0.000); the average durational values are ms.56.4 (V2) and ms. 111.54 (V3). The increment in duration of the final vowel is greater in unstressed than in stressed syllables (see, Fig.7, stressed V vs unstressed V in Isolated Words).
Also vowel posteriority increases significantly in syll.3 vs syll.2 in isolated and final words; the average PI values are PI=4.34 for V2, PI= 5.77 for V3. The increment in PI is more consistent in unstressed than in stressed vowels. For GA it is significant in both stress conditions (p=.04 0.018 for Str.V; p=0.000 for Uns.V;), for EF only for unstressed vowels (p=.04 0.003). As can be seen in Fig.7, the changes in PI and in VDUR in final vowels are similar to those induced by initial boundaries, but are much larger for both parameters. It has to be noted that in embedded words, where the boundary is not realized, the duration of the word final vowel tends to be even shorter than the word medial vowel (see Fig.7, EM, Subj. EF).

3.2.5. Comparing the articulation of stress and boundaries

To a certain extent, the articulation of boundaries resembles the articulation stress: duration and front contact increase in consonants, posteriority increases in vowels, accompanied by a relevant increase in duration in absolutely final vowels. These combined effects on C and V make initial and final unstressed syllables more prominent than word-medial unstressed ones, while a stressed syllable at the word edges further enhances the prominence contrast at the word level. But there are important differences between the effects of stress and those of boundaries, first, quantitative differences: in vowels, the spatiotemporal changes induced by boundaries are smaller than those induced by stress (compare Fig. 2 and Fig.3 with Fig.7), and, second, stress induces also an increase in centrality in high vowels, which indicates that the tongue body has a higher position and a tighter constriction. Boundaries do not induce such effects, on the contrary, when CI changes, it tends to decrease at the boundaries. Thus, the articulation of boundaries comports mainly a larger opening of vowels and a

tighter closing of consonants, whose outcome is an enhancement of the V/C contrast. Instead, the articulation of stress, comports not only an expansion of vowel opening and consonant closing movements, but also an increase in the target-directed movement for each vowel, hence, an enhancement of the vowel specific distinctive properties and a consequent increase in the phonetic distance among the different vowels.
As for vowel duration, while final boundaries induce an increase in duration of both stressed and unstresssed final vowels, with stronger effects on unstressed vowels (as can be seen in Fig.7 for isolated words), initial boundaries induce a lengthening of unstressed vowels only (see Fig.7 all contexts). The stressed vowels of syll.1 ('CVCVCV words) are never longer than those in syll.2 (CV'CVCV words), on the contrary, they are significantly shorter in isolated words (see Fig.7, IS). These data on stressed vowel duration agree with a previous study on duration as a function of word size and stress position (Farnetani and Kori, 1986). In that study the data indicated that stressed vowels are longer in 'CVCV words than in 'CVCVCV but not in CV'CV vs CVCV'CV words, which showed that the durational differences were not related to word size but rather to the number of unstressed syllables following the stressed one, in other words the stressed Vs shorten as the word extends rightwards. Thus, in the present study, the shorter duration of stressed vowels in the initial syllables of 'CVCVCV words vs the medial syllables of CV'CVCV words are to be related to the greater distance of the stressed syllable from the end of the word, and can be accounted for by rhythmical constraints.

4. General discussion and conclusion

The first purpose of the present research was to know to what extent the prosodic features of lexical stress and boundaries influence the articulation of a segment. The data indicate that the articulatory spatial changes are systematic, quantitatively relevant and that in most cases they are not mere consequences of durational changes, but rather active articulatory modifications aiming at the production of various degree of articulatory prominence. As seen before, in the production of stress and boundaries both the consonant and the vowel of the CV syllable play an important role. As for the effects of the prosodic context, the data have shown that lexical prominence is not weakened when duration is compressed in words embedded within a prosodic phrase, and that in phrase internal position the word initial boundaries are preserved, although weakened, while word final boundaries are totally cancelled. Thus it appears that the syllabic prominence induced by lexical stress and, to a lesser extent, the prominence resulting from the articulation of initial boundaries are maintained also at the word level.
The second purpose of this study was confront the present data with the models proposed for the production of prominence and boundaries. The inferences that can be drawn from the analysis of the spatial configuration of C and V are that lexical prominence, initial boundaries, and final boundaries, though similar in their articulation under certain respects, differs substantially among each other: it has been shown that the articulation of stressed syllables is characterized by a strengthening of C, an expansion of vowel opening and an enhancement of vowel specific articulatory features, and this would agree with the hypothesis that a stressed syllable is a hyperarticulated syllable. This cannot be said, at least for high vowels, for the articulation of boundaries: both initial and final boundaries do induce induce a strengthening in the consonant but only an expansion of opening in the vowels. Moreover there are substantial differences between initial and final boundaries: initial boundaries are marked articulatorily also when duration does not increase, while final boundaries are marked more strongly by duration than by an increase in articulatory prominence.
In order to gain some more knowledge on the control strategies underlying the production of stress and boundaries, multiple regressions have been carried out for vowel /a/. Even if tongue movements cannot be recorded with EPG, they can be indirectly inferred from the traces left by the tongue-to-palate contact over time. In this analysis the number of free electrodes in vowel /a/, measured in the first temporal frame of minimum contact was the dependent variable of a number of stepwise regressions with the amount and the duration of the C-to-V displacement, the mean displacement rate and the maximum (or peak) displacement rate were the independent variables. The purpose was to infer how prominence, initial boundaries and final boundaries are produced in terms of C-to-V gestures. The characteristics of C-to-V gestures for stress were analyzed in word medial syllables, those for initial and final boundaries were analyzed by confronting word medial syllables with word initial and word final syllables, respectively. The preliminary results indicate that the very low contact characterizing stressed /a/ is in general accounted for by an increased duration and displacement of the C-V movements. But in

embedded words stress is associated with the duration and the peak rate of the C-V movement. Initial boundaries production (unstressed syllables) is associated with the duration and the rate of the movements for GA and to the their peak rate for EF, while the articulation of final boundaries (unstressed syllables) seems to be accounted for only by the duration of the C-V movements for GA and by displacement and duration for EF. This preliminary data, in which there is a good agreement between the two subjects, suggest that hyperarticulation characterizing stressed vowels may be simply the outcome of a sufficient time allowed to the articulators to achieve the target; however, in embedded words where segments undergo temporal compression the stressed vowel is still hyperarticulated, and in this case the target is achieved by increasing the rate of the gestures. The two different articulatory strategies inferred from the present data are compatible with the revised H&H model (Moon and Lindblom, 1994) were *duration* and *input force* are two of the predictors of vowel undershoot and reflects the strategies available to speakers under different circumstances. Hence, the sonority expansion hypothesis (Beckman et al.1992) according to which prominence is brought about by decreasing the temporal overlap between adjacent gestures, can account only in part for the present data. As for the articulation of the vowel in the first syllable of the word, where the spatial changes are larger than the temporal ones, the articulatory movements must be faster in order to attain a greater opening of the vocal tract in the production of an unstressed vowel without interfering with the rhythm imposed by the stressed/unstressed contrast. Final boundaries seem to be articulated by slowing down the speech tempo (which could account for the very long vowel duration) and/or by expansion of the jaw movements (which could account for a an increase in contact for the consonants and an increase in opening for the vowels). Further analysis is now in progress for high vowels.

References

Beckman, M.E., Edwards, J.R., and Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. Docherty and D.R. Ladd (eds), *Papers in Laboratory Phonology II, Segment, Gesture, and Tone*. CUP, pp. 68-86.

Browman, C.P., and Goldstein, L. (1989). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M.E. Beckman (eds), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech.* CUP, pp. 341-376.

de Jong, K.J. (1995). The supraglottal articulation of prominence in English. Linguistic stress as localized hyperarticulation. JASA 97, 491-504.

Edwards, J.R., Beckman, M.E., and Fletcher, J. (1991). The articulatory kinematics of final lengthening. JASA 89, 369-382.

Farnetani, E. (1989). Acoustic correlates of linguistic boundaries in Italian: A study on duration and fundamental frequency. Proceed. of European Conference on Speech Communication and Technology. Eurospeech 89, Vol.2, 332-335.

Farnetani, E., and Kori, S.(1983). Interaction of syntactic structure and rhytmical constraints in the realization of word prosody. Quaderni CSRF 2, 288-318.

Farnetani, E. and Kori S. (1986). Effects of syllable and word structure on segmental durations in spoken Italian. Speech Communication 5, 17-34.

Farnetani, E. and Kori, S. (1990). Rhythmic structure in Italian noun-phrases: A study on vowel duration. Phonetica 47, 50-65.

Farnetani E. and Faber A. (1992). Tongue-jaw coordination in vowel production: Isolated words vs connected speech. Speech Communication 11, 401-410.

Farnetani, E. and Vayra, M. (1996). The role of prosody in the shaping of articulation in Italian CV syllables. Proceed. 1st ESCA Tutorial and Research Workshop on Speech Production Modeling. Autrans, 9-12.

Fougeron, C. and Keating, P. (1996). The influence of prosodic position on velic and lingual articulation in French: Evidence from EPG and airflow data. Proceed. 1st ESCA Tutorial and Research Workshop on Speech Production Modeling, Autrans, 93-96.

Keating, P. (1995) Effects of prosodic position on /t,d/ tongue/palate contact. Proceed. ICPhS 95, Stockholm, V3, 432-435

Keating, P. (in press). Word-level phonetic effects on English consonant articulation . Paper presented at the Conference on the Phonological Word. ZAS, Berlin, October 1997.

Kohler, K.J. (1990). Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In W.J. Hardcastle and A. Marchal (eds), *Speech production and speech modeling*. Dordrecht: Kluwer Academic Publishers, pp. 69-92.

Lea, W.A. Prosodic aids to speech recognition (1980). In W.A.Lea (ed) *Trends in Speech recognition*, Prentice Hall Inc. New Jersey, pp. 166-205.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W.J. Hardcastle and A. Marchal (eds), *Speech Production and Speech Modelling*, Kluwer Academic Publishers, pp. 403-439.

Lindblom, B. and Repp, K. (1973). Some temporal regularities of spoken Swedish. Papers from the Institute of Linguistics, Stockholm University, 21, 1973, Monograph

Macchi, M. (1985).Segmental and suprasegmental features and lip and jaw articulators. Ph.D. dissertation, New York University.

Moon, S.J. and Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. JASA 96, 40-55.

Perkell, J.S. (1997). Articulatory Processes. In W.J.Hardcastle and J.Laver (eds), *The Handbook of Phonetic Sciences,* Blackwell Publishers, pp.333-370.

Word Boundary Marking at the Glottal Level in the Production of German Obstruents*

Michael Jessen
University of Stuttgart

1. Introduction

Despite the impression – nourished by the regular use of written language among the literate community – that an utterance can easily be segmented into words, phonetic research has shown that on the signal-phonetic level of reality it is very difficult to observe and understand how the unity and demarcation of words is manifested. When, for example, the spectrogram of a portion of naturally produced speech is evaluated in a naïve way by expecting pauses between words, it is quite likely that the portions of silence or near-silence created by the closure phases of stops (especially when voiceless) are more salient than whatever might exist in terms of word demarcation, and that consequently stop closures would be mistaken for word boundaries (thanks to Anne Cutler for a demonstration of this effect). In the search for cues to word segmentation in the phonetic signal it is not known a priori on which location in the syntagmatic dimension, on which level of the speech chain, or with which choice of signal-processing algorithms this search is most likely to succeed. Most of it is a matter of systematic empirical phonetic research on all possible levels.

The present contribution attempts to provide another piece in the puzzle of approaching the word as a phonetic unit. The phonetic domain that will be investigated is speech production at the level of gestures. Based on the results of a transillumination study, we will evaluate patterns of the glottal opening gesture in the voiceless obstruents of German under the production of sequences with and without an intervening word boundary. After reviewing the literature about the potential of glottal opening gestures for the coding of word segmentation, this study will show that the facts provided here are better explained with an acoustic-based than a gesture-based interpretation of word segmentation.

2. Results and perspectives from the literature

Transillumination provides a valid and reliable means of visualizing and quantifying the glottal opening movements that are associated with voiceless stops and fricatives or with /h/ (see Hoole, to appear, for overview). Among the different possible experimental phonetic agendas that can be pursued with transillumination, one major subject of interest has been the glottal opening patterns in sequences of voiceless obstruents and the influence that a word boundary or related morpho-syntactic boundary within these obstruent sequences has on glottal opening behavior.

Frøkjær-Jensen, Ludvigsen & Rischel (1971) found that if a sequence of /s/ followed by /p/ in Danish is separated by a word boundary and if it is produced relatively slowly, the transillumination signal indicates that instead of the single glottal opening gesture present in individual voiceless obstruents a compound glottal opening occurs that is suggestive of two underlying/intended gestures, one for /s/ and one for /p/ (which is aspirated word-initially in Danish). Frøkjær-Jensen, Ludvigsen & Rischel mention the possibility that this, what they call two-peakedness, is a boundary/junctural phenomenon. They also show, however, that as soon as the same sequence is produced less slowly with what they refer to as normal pronunciation the effect disappears and only a one-peak glottal gesture occurs. For the remainder of this contribution a one-peak glottal opening pattern will be referred to as "monomodal", and a two-peak pattern as "bimodal".

In contrast to the Danish case of /s#p/ (# symbolizing a word boundary), in which a compound glottal opening is at least potentially possible, Lindqvist (1972) found for Swedish that

(Löfqvist & Yoshioka 1981), and Dutch (Yoshioka, Löfqvist & Collier 1982). Most of the discoveries and conclusions relevant for the word boundary marking issue were already made in the Löfqvist (1978) and Löfqvist & Yoshioka (1980) studies on Swedish. They were later confirmed for the other Germanic languages as well (cf. also Fukui & Hirose 1983 on Danish and Hoole, Pompino-Marschall & Dames 1984 on German for further consistent evidence). Among all the interesting details reported in the work of Löfqvist and colleagues we concentrate on those cases that fall into the falsification patterns mentioned in (2).

One regular exception to (1) is the case of adjacent identical obstruents.[1] Even if separated by a word boundary only a monomodal glottal opening is found over adjacent identical obstruents. For example, the sequences /s#s/ in *My **ace** **sales*** and /k#k/ in *I **make** **cave*** are produced with a single glottal opening (Yoshioka, Löfqvist & Hirose 1981). This effect also occurs if certain other obstruents precede or follow the sequence of adjacent obstruents. This situation applies to the sequence /ks#sp/ of Swedish ***Eks** **spelar*** 'the Eks play', which is produced with a single glottal opening (Löfqvist & Yoshioka 1980: 798). The case of monomodal glottal opening gestures in identical obstruents that are separated by a word boundary constitutes a type (2a) falsification of the Word Boundary Marking Hypothesis in (1).

The data provided by Löfqvist and colleagues also contain cases that constitute type (2b) violations of the Word Boundary Marking Hypothesis. Löfqvist & Yoshioka (1980: 797) show for Swedish that the sequence /sts#/ in ***Kvists** ilar* 'the Kvists hurry' is produced with a bimodal glottal opening. This is the case although the obstruent cluster is not interrupted by a word boundary, but only followed by it. Apparently, the explanation for the compound glottal opening in this case cannot be found in terms of word segmentation, but it must be a matter of the specific obstruents involved and their sequencing pattern. Another example that illustrates the same phenomenon is the trimodal glottal opening pattern found in the sequence /sts#p/ of Swedish ***Kvists** **pilar*** 'the Kvists throw'. In this case the word boundary can explain at most two of the three peaks in the sequence, whereas the two peaks associated with the /sts/-part of the sequence must be due to other factors, as in the previous case.

To explain the entire range of glottal opening patterns Löfqvist & Yoshioka (1980: 800) write: "The observed variations in glottal area are obviously related to the segmental properties of an utterance. The most apparent relation is that sounds requiring a high rate of air flow are produced with a separate glottal opening gesture". Analogous statements are made in Löfqvist (1978) and various of the work with his colleagues, mentioned above. In (3) this generalization is repeated with a slightly different wording.

(3) *Löfqvist's rule*:

> A separate glottal opening peak is associated with every aspirated stop and every voiceless fricative, in order to fulfill the necessary aerodynamic conditions for the production of these sounds.

If (3) turns out to be correct (it will be argued that it does), word boundary marking is no longer an immediate explanation for the articulatory patterns of laryngeal gesture organization. Word boundaries, of course, have to be manifested somehow. It seems that as far as voiceless stops are involved this phonetic coding of word boundaries is most straightforwardly provided by the presence or absence (or the absolute degree of) *aspiration*. Consider again the case investigated by Pétursson (1977), in which the sequence /st/ was produced with a single glottal opening word-internally, but a bimodal glottal opening if separated by a word boundary (analogous cases exist for other Germanic languages). The crucial point is that in the Germanic languages investigated by Pétursson, as well as Löfqvist and colleagues, voiceless stops are aspirated when they occur word-initially without further obstruents, but that they are unaspirated when occurring with a preceding

voiceless fricative in word-initial position. In the pattern /s#t/ word-initial /t/ is aspirated and hence receives a glottal opening peak of its own, while the word-final voiceless fricative /s/ also receives its own glottal peak according to (3). In the pattern /#st/, on the other hand, the /t/ is unaspirated and receives no glottal peak of its own. Instead, it shares one single glottal opening with the one that already exists due to the voiceless fricative /s/.

Beyond the case of /s#t/ versus /#st/, which could also be explained by the Word Boundary Marking Hypothesis in (1), Löfqvist's generalization in (3) can also capture the problematic cases of the types (2a) and (2b). That with faster speech rates and less formal speech a bimodal gesture in a structure like /s#t/ can be transformed into a monomodal pattern might be explained by the reduction of aspiration duration that is expected to occur with these stylistic changes (cf. Löfqvist 1992: 19). And the bimodal gesture in the word-internal sequence /sts/ found by Löfqvist & Yoshioka (1980) can be explained by the fact that this sequence contains two voiceless fricatives and one unaspirated stop, which according to (3) amounts to a bimodal pattern. In the sequence /sts#p/ with the trimodal pattern found by Löfqvist & Yoshioka the third peak is due to the free-standing word-initial aspirated /p/. The only case that remains problematic even under the account in (3) is the monomodal pattern of adjacent identical obstruents, in particular fricatives. In the sequence /s#s/ two glottal opening peaks are expected, although as a rule only one peak is found (Yoshioka, Löfqvist & Hirose 1981). Contrary to the fricative sequence, a sequence of identical stops across a word boundary is expected to be realized with a monomodal glottal opening if the word-final stop is unaspirated (which is often the case in English). This prediction is borne out for the sequence /k#k/ shown by Yoshioka, Löfqvist & Hirose (1981), as well as the sequence /p#p/ shown by Lisker & Baer (1984: 167). Despite the unexpected monomodal pattern in the fricative sequence there is still evidence in the EMG data of Yoshioka, Löfqvist & Hirose (1981) for a compound structure in /s#s/ as opposed to a truly plain structure in /k#k/ (the same asymmetry between fricative geminates and stop geminates is reported for Japanese by Yoshioka, Löfqvist & Hirose 1982). The merging of glottal gestures in sequences of identical obstruents could be due to a general degemination tendency, that is known for many languages, especially in systems like English, German etc., which do not tolerate true geminates. Also problematic is the finding mentioned by Yoshioka, Löfqvist & Collier (1982) that a monomodal glottal opening pattern is associated with word-initial /sx/ in Dutch, although these two voiceless fricatives are not identical.

This discussion of the literature has shown that the Word Boundary Marking Hypothesis in (1) has been met with skepticism by most of the phoneticians that performed the relevant trans-illumination (and related) experiments. One could argue that due to this skepticism and the success of the alternative explanation in (3) it would be of no current relevance to pursue this issue any longer. However, it seems that among a number of phonologists (1) is still assumed in one or the other fashion. We will concentrate on two proposals along these lines, one made by Browman & Goldstein (1986) and one by Iverson & Salmons (1995).

We saw that in the word-initial voiceless fricative-stop clusters of languages like English and Swedish the voiceless stop is unaspirated (e.g. in a word like *spill*). In order to provide a plausible phonological explanation of this fact, Browman & Goldstein (1986: 227) propose that in the case of clusters such as word-initial /sp/ a single glottal opening gesture coincides with several supraglottal gestures, in this case one alveolar fricative gesture for /s/ and one bilabial closure gesture for /p/. Since the single glottal gesture is approximately coordinated with the fricative, both if occurring individually or if followed by a stop (this is shown in more detail in the literature of Löfqvist and colleagues), there is not sufficient time left after the release of a stop in clusters like /sp/ during which the glottis would be open and thus in an appropriate condition for the production of aspiration. Browman & Goldstein claim that this suprasegmental status of glottal opening, by which the supraglottal gestures of more than one segment are associated with one shared glottal opening gesture, is a typical feature of word-initial consonant clusters in English and other Germanic languages. Browman & Goldstein's position is expressed in (4).

(4) *Browman & Goldstein's rule*:

Words in English and other Germanic languages begin with at most a single glottal gesture.

While the generalization in (4) has no implications for the (2a) problem, is makes the prediction that the false alarm problem in (2b) should not occur, at least not word-initially in Germanic languages. The bimodal glottal opening observed by Löfqvist & Yoshioka (1980) for the word-final sequence /sts/ in Swedish, although being a (2b) type counterexample of (1), is not strictly a counterexample of (4), simply because it occurs word-finally (in the Swedish name *Kvists*) and not at the beginning of the word. If, on the other hand a case should be found in which any sequence of consonants is produced with more than one single glottal opening gesture *word-initially*, Browman & Goldstein's generalization in (4) would be violated (two cases of this type will be shown for German in § 4).

Iverson & Salmons (1995) discuss the same facts that were also brought up by Browman & Goldstein (1986). They provide similar arguments, but arrive at a proposal that differs from the use of articulatory gestures, that is characteristic of Browman & Goldstein's approach to phonology as a whole (see Browman & Goldstein 1992). Working broadly within the tradition of nonlinear phonology (see Kenstowicz 1994), Iverson & Salmons represent glottal opening in the production of voiceless obstruents with the phonological feature [spread glottis], which goes back to a proposal by Halle & Stevens (1971). The case of word-initial clusters of the type /sp/ is represented by Iverson & Salmons with a single instance of the feature [spread glottis] that is associated with the feature matrices of both the fricative and the following stop. With this representation they intend to capture essentially the same as Browman & Goldstein did, namely that only a single glottal opening is associated with the entire /sp/-like cluster and that this prevents the stop from being aspirated. But aside from the broader theoretical implications there is one important difference between the two approaches that has empirical consequences. By considering the multiple association of [spread glottis] as a special case of the Obligatory Contour Principle (see McCarthy 1988) Iverson & Salmons (1995) claim that this multiple association pattern occurs in any word-internal (or, more accurately, morpheme-internal) position, including word-initial, word-medial, and word-final position. In a format similar to (4) this claim is stated in (5).

(5) *Iverson & Salmon's rule*:

At most a single glottal gesture is found for any word-internal consonant sequence in English and other Germanic languages.

The statement in (5) makes the empirical prediction that more-than-single-peak glottal openings should not occur for any word-internal consonant sequence. But this prediction is too strong if we again think about the case of word-final /sts/ in Swedish with its bimodal glottal opening. It might be that with further formal assumptions this case might be accommodated to the analysis of Iverson & Salmons (1995), but as it stands now, evidence of the type present in the Swedish /sts/ cluster is problematic to their account. In order to provide more data that are relevant to the issues discussed here the results of a transillumination study of German will be illustrated and discussed for the remainder of this contribution.

3. Methods

In the transillumination procedure used in this study the glottis is illuminated with the help of a fiberscope that is inserted through the nasal cavity. The light source is attached from outside and feeds cold light of sufficient brightness into the fiberscope. The tip of the fiberscope is positioned

in the hypopharynx in a way which provides an unobstructed view of the glottis. The amount of light that passes first through the glottis and then through the tissue of the neck skin is measured with a phototransistor that is attached to the neck and held in place with the help of a neckband. The phototransistor was held at the level of the cricothyroid membrane. The subject (the author) was seated in a dental chair. The fiberscope was inserted through the nasal cavity by an otolaryngologist (Kiyoshi Oshima) under the application of some local anesthesia. Insertion and positioning of the fiberscope was controlled by the physician with the help of the image from a standard video system. The stimulus words were printed in large letters, randomized, and placed on a stand that was adjusted in height to ensure that the list was visible to the subject. Recordings were made both of the transillumination signal and of the audio signal. The audio signal was recorded with a Sennheiser direction-sensitive microphone that was positioned in front of the subject. A number of test tokens were produced by the subject, during which the levels of the transillumination and the audio signal were controlled by the director of the experiment (Anders Löfqvist) with the help of an oscilloscope. The input volume was adjusted such that the audio signal was unclipped for the relevant portions associated with obstruent production. The recording took place at Haskins Laboratories. Both the audio and the transillumination signal were directly digitized into the computer with a sampling rate of 10 kHz, each using preemphasis. The transillumination signal was smoothed using a triangular window of size 35.1 ms. This eliminates the voicing patterns of the signal and facilitates the evaluation of the glottal opening patterns associated with obstruent production.

The primary purpose of the experiment was an investigation of the difference between voiced and voiceless obstruents in German, as documented in Jessen (to appear). Due to this emphasis most of the material read by the speaker contains obstruents in contexts in which they contrast in voicing. This is the case for the data discussed in § 5, where the obstruents occur word-initially and are preceded by a voiceless palatoalveolar fricative across a word boundary. Relevant for the present study are only the voiceless target obstruents. The different patterns investigated for the discussion in § 5 are /ʃ#p/, /ʃ#t/, /ʃ#k/, /ʃ#f/, and /ʃ#s/ (due to the primary interest in the voicing opposition no fricatives at other than labial and alveolar place of articulation were included, since others do not engage in the voiced/voiceless opposition). The patterns were created by combining the carrier word *rasch* /ʃ/ 'quickly' with the words ***Pier*** /p/ 'pier', ***Tier*** /t/ 'animal', ***Kir*** /k/ 'kir', *vier* /f/ 'four', and ***Sir*** /s/ 'sir', respectively. Some of these examples are somewhat marginal, but they have the advantage of constituting minimal pairs and of containing a following vowel /i/, which is the preferred vowel for transillumination studies. The carrier word *rasch* was preferred over other possible carriers ending in one of the target obstruents, in order to avoid artefacts that might occur with identical adjacent obstruents (see § 2). Each target pattern was repeated approximately twenty times in blocks of approximately ten. The same number of tokens were produced in a second session three weeks later. The exact positioning of the fiberscope in a recording session, as well as the exact placement of the phototransistor, have effects on the intensity level and other details of the transillumination trace. For this reason the results for the two sessions are kept apart in the evaluation of the results. Contrary to § 5, the material discussed in § 4 was less substantial. It contains voiceless obstruents in word-initial clusters such as /ʃp/, /sk/, /ʃpl/ etc. (in this position no voiced/voiceless opposition is possible in German). The material was only repeated a few times and only recorded in one of the two sessions for exploratory purposes. Yet, this material should still be important and interesting enough for some basic conclusions and ideas about what might be expected in a more in-depth investigation of these patterns. Further information on the methods used here is found in Jessen (to appear).

4. Multiple word-internal glottal openings

In this section some of the glottal opening patterns will be discussed that occur in word-initial obstruent clusters in German. Special emphasis will be given to those cases that are immediately

relevant for the different hypothesis in the literature that were discussed in § 2. Figure 1 provides an illustration of four subsequent repetitions of the sequence /i#ʃpi/ in the utterance *nie Spier* 'never spar'. The top part of Figure 1 shows the audio waveform, the lower part the smoothed transillumination signal. Both representations are temporally aligned and labeled at the bottom for the occurring sequences of sounds. The transillumination curve shows that the word-initial sequence /ʃp/ is produced with a bimodal glottal opening in the first, third, and fourth repetition. No clear two peaks are observable in the second repetition, but the glottal opening gesture still looks complex. This bimodal pattern for word-initial /ʃp/ is in violation of the claims of both Browman &

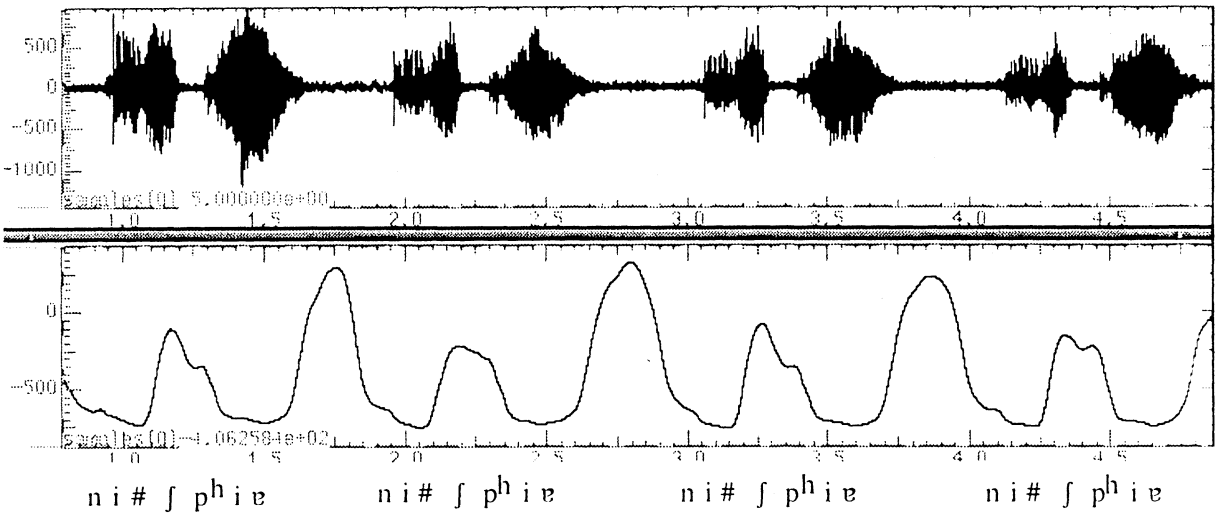


Figure 1: Bimodal glottal opening in /#ʃp/.

Goldstein (1986) in (4) and Iverson & Salmons (1995) in (5). Is it also in violation of Löfqvist s generalization in (3)? Measurements of aspiration duration that were carried out on the basis of spectrograms derived from the acoustic signal indicate values that on an average almost reach 50 ms.[2] Although it is difficult and controversial to divide a continuum of aspiration duration values into the categories aspirated vs. unaspirated, values of almost 50 ms allow the conclusion that at least some aspiration was present in these stops. Apparently the common assumption that stops after word-initial fricatives are necessarily unaspirated in Germanic languages is too strong. Lotzmann (1975) in his transcription-based study of aspiration in German provides a discussion of the literature, including sources that address dialectal differences. He shows that in North German pronunciation aspiration of /p,t,k/ is more common and stronger than in the pronunciation of South German speakers (pp. 9-31). This is consistent with the fact that the speaker in the present study was raised in North Germany (close to Flensburg). He also shows that in the corpus of speech that he investigated stops in the word-initial clusters /ʃp/ and /ʃt/ can be aspirated, even if this is a less preferred option. He found the stop in /ʃp/ to be 34 times aspirated and 101 times unaspirated and the one in /ʃt/ to be 39 times aspirated as against 213 times unaspirated (p. 135). Knetschke & Sperlbaum (1987: 149) in their large transcription study mention that stops in fricative-stop clusters like /ʃp/ and /ʃt/ were aspirated "relatively often" in the speech of German newscasters. These results indicate that aspiration in word-initial fricative-stop clusters is an option German, which had been implemented in the tokens shown in Figure 1. Given that the stops in Figure 1 are aspirated, the finding of a bimodal glottal opening is not in violation, but in confirmation of Löfqvist's generalization in (3).

Another set of examples relevant for the present discussion is shown in Figure 2. The transillumination signal above shows four subsequent repetitions of the sequence /i#ʃpli/ in the utterance *nie Splier* and the one below shows four repetitions of the sequence /i#ʃpri/ in the

utterance *nie Sprier* (*Sprier* and *Splier* are nonsense words, but phonotactically normal in German). It can be seen in Figure 2 that across repetitions a single glottal opening occurs for the sequence /ʃpl/, but that a bimodal glottal opening occurs for /ʃpr/. To understand these patterns it is important to point out that /r/ was produced as a voiceless uvular fricative [ʁ̥] in this context. The realization of German /r/ as an uvular fricative is the preferred realization of this phoneme in the syllable onset, and its manifestation as fully voiceless after other voiceless obstruents in the syllable onset is common for German as well (Kohler 1995: 166). If we take into account that the

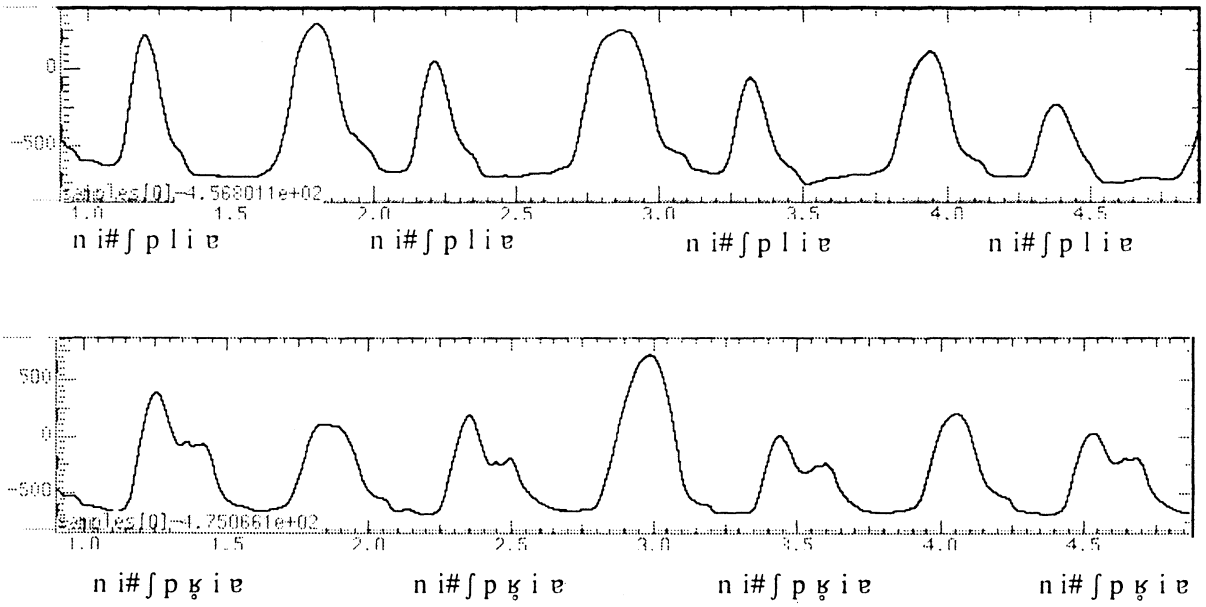


Figure 2: Monomodal glottal opening in /#ʃpl/, bimodal in /#ʃpr/.

sound /r/ in /ʃpr/ is realized as a voiceless fricative, the bimodal glottal opening follows from Löfqvist's generalization in (3). Since the sequence /ʃpr/ contains two voiceless fricatives and one unaspirated stop, two glottal opening peaks should occur, which is borne out in Figure 2. In contrast, the /l/ in /ʃpl/ bears no resemblance to a voiceless fricative and the stop is unaspirated; hence only the single glottal opening associated with /ʃ/ should be found, which is again borne out. The lack of aspiration of the stops here (as measurable in the audio signal, which is not shown here) as opposed to the aspiration of the stops in Figure 1 might be partially explained by the universal phonetic tendency that the duration of segments shortens with the number of segments in a cluster (see Klatt 1976).

Taken together, Figures 1 and 2 have shown us two cases of glottal opening behavior that are explainable by the generalization of Löfqvist in (3), but that present direct counterevidence to the phonological hypotheses in (4) and (5). These two cases show additional instances of the false alarm problem of the Word Boundary Marking Hypothesis in (1), which add to the bimodal pattern found by Löfqvist & Yoshioka (1980) in the word-final sequence /sts/ of Swedish.

5. Gestural aggregation

We now turn to cases that challenge the Word Boundary Marking Hypothesis from the direction of the problem of misses in (2a). It will be shown that a monomodal glottal opening is possible despite the presence of an intervening word boundary. The most systematic research along these lines has been performed by Munhall & Löfqvist (1992). Munhall & Löfqvist show that the sequence /s#t/ is produced with two separate glottal openings in slow speech. With a gradual increase in speech rate the two separate glottal openings turn into a compound bimodal structure, until they finally merge into a plain monomodal glottal opening. This merging of glottal openings

with increasing rate of speech is referred to as "gestural aggregation" by Munhall & Löfqvist. The German data to be discussed in this section show that a substantial degree of gestural aggregation occurs even in the relatively slow rate of lab speech that was used in the present study, and that factors other than rate of speech must be held responsible as well.

A corpus of about 200 tokens (ca. 20 tokens • 2 sessions • 5 word initial voiceless obstruents; cf. § 3) was evaluated qualitatively. This was done by printing out each one of the tokens, arranging them according to session number as well as place and manner of articulation of the second obstruent, and trying to find generalizations by "eye balling" the material. One striking observation was the considerable token-to-token variability of the same sequence repeated successively within the same session. It was not at all uncommon to find a very clear bimodal structure being followed in the next repetition by a glottal opening with only a minimal trace of compoundedness left. Fortunately, aside from this high degree of pure free variation it was also possible to observe some linguistically systematic tendencies. As will be shown and explained for the remainder of this section, it was found that gestural aggregation depends to a certain degree on the place and manner of the word-initial obstruent in the position after word-final /ʃ/. As a tendency, the degree of gestural aggregation occurs in the order labial > alveolar > velar and in the order fricative > stop. The representative examples in Figure 3 illustrate these tendencies. The transillumination signal at the top of Figure 3 shows an example of the sequence /ʃ#f/ in *rasch vier* 'quickly four'. The vertical scale of glottal opening intensity is in arbitrary units and is adjusted to fit the entire range of the window, which is the reason why different ranges of values occur in the vertical axes of the three examples (the same holds for any other transillumination signal shown in the figures of this paper). It appears that the glottal gesture visible here is a bit more complex than would be expected from a plain single glottal opening, but it is impossible to see more than one peak. The signal in the middle of Figure 3, which shows an example of the sequence /ʃ#p/ in *rasch*

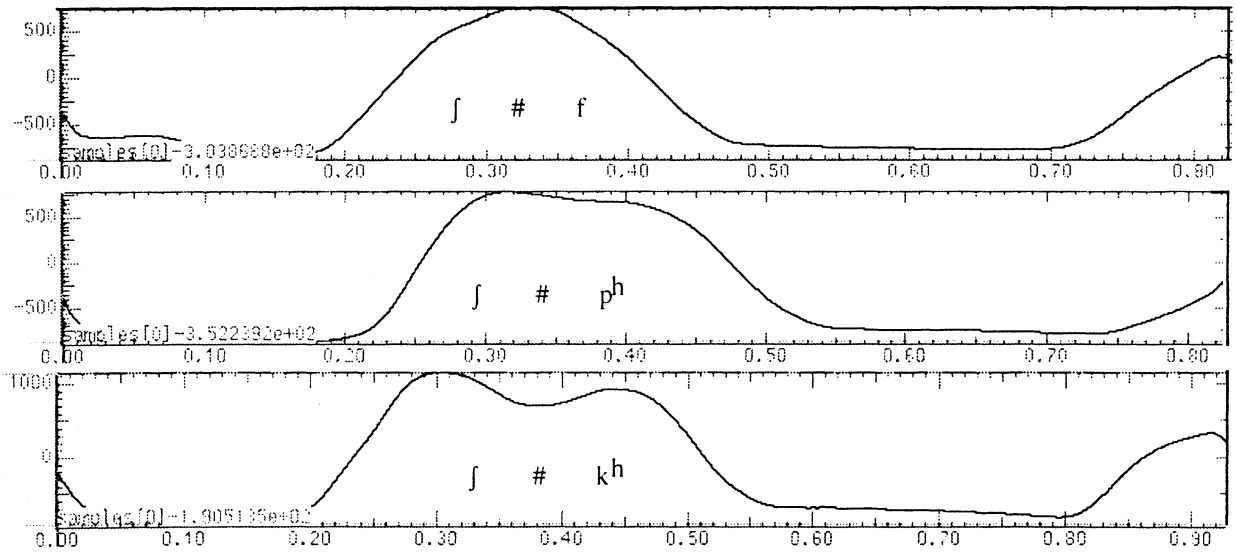


Figure 3: Place- and manner-related differences in gestural aggregation.

Pier emerges as clearly complex, but it is difficult to tell apart the dimension of the first from that of the second component. Finally, at the bottom of Figure 3 it is shown that in the sequence /ʃ#k/ in *rasch Kir* a bimodal glottal opening can be determined with the two peaks clearly distinguishable. From the top to the middle part the influence of manner of articulation can be seen (fricative vs. stop, respectively), while from the middle to the bottom part the influence of place of articulation (here: labial vs. velar, respectively) is identifiable. In each of these steps the compound nature and bimodal appearance becomes clearer.

An explanation for the situation illustrated in Figure 3 can be sought in the different oral-laryngeal timing patterns of voiceless obstruents that differ with respect to place of articulation and the stop-fricative distinction. In the investigation described more explicitly in Jessen (1995) and Jessen (to appear) the dimension and oral-laryngeal coordination of glottal opening has been measured in word-initial obstruents preceded by a vowel. The target words were the same as the ones described in this section (*Pier, Tier, Kir, vier, Sir*), with the only difference that the preceding carrier word ended in a vowel (in the word *nie* /ni/ 'never', as in *nie Pier, nie Tier* etc.) instead of the fricative /ʃ/ used in the present study. Since in those stimuli the target obstruents were surrounded by vowels it was much easier to quantify the entire glottal gesture, whereas in the present setting full quantification was inhibited due to the influence of the glottal gesture associated with /ʃ/. Table 1a,b shows a selection of those results in the investigation of Jessen (1995) and Jessen (to appear) that reveal differences in oral-laryngeal timing that are due to place (Table 1a) or manner (Table 1b) of articulation. The particular parameters addressed in the table are the interval between the onset of the target obstruent and the onset of its glottal opening (OG-OC) and the interval between the onset of the target obstruent and the moment of peak glottal opening associated with that obstruent (P-OC). Table 1a shows the means, as well as the standard deviations (in parentheses), of the parameters OG-OC and P-OC (in ms) for labial, alveolar, and velar place of articulation and the P-values that emerge in one-factor ANOVAs with place of articulation as the independent variable and OG-OC, as well as P-OC as the dependent variables (see different columns). In the case of stops, in which three places exist, the probabilities for each pairwise comparison (Tukey HSD multiple comparisons) is provided separately (with l = labial, a = alveolar, and v = velar). Separate analyses were carried out for stops in the first recording session, stops in the second session, fricatives in the first, and fricatives in the second session (see different rows). Each of the

Table 1a: Oral-laryngeal coordination for different places of articulation.

		OG-OC				P-OC			
Manner	Sess	labial	alveolar	velar	P	labial	alveolar	velar	P
stops	1	5 (7)	5 (3)	6 (11)	l/a: .96 l/v: .90 a/v: .98	125(10)	124 (7)	130(11)	l/a: .86 l/v: .28 a/v: .11
	2	7(5)	16 (10)	17 (19)	l/a: .10 l/v: .07 a/v: .98	111 (7)	115 (7)	128(14)	l/a: .55 l/v: .00 a/v: .00
fricatives	1	-9 (4)	-14 (9)	-	.07	104 (7)	111(15)	-	.1
	2	-14 (8)	-6 (6)	-	.001	95 (9)	105(13)	-	.007

mean values are based on approximately 20 repetitions. Table 1a shows that as a tendency OG-OC and P-OC increase as place of articulation proceeds backwards in the oral cavity. This tendency holds for both stops and fricatives. Although consistent in most cases, this effect is only significant in some cases, i.e. the order velar > labial and velar > alveolar in the parameter P-OC for stops produced in the second recording and the order alveolar > labial in both parameters for fricatives of the second recording.[3] Table 1b shows the results of a comparison between stops and fricatives (columns), separately for labial place of articulation in the first and second, and alveolar place of articulation in the first and second recording (rows). Beyond this difference, the conventions used for Table 1b are identical to Table 1a. Table 1b shows quite a robust pattern according to which higher values of OG-OC and P-OC are found for stops than fricatives. This effect holds across the two different places of articulation involved, and it is statistically significant in each case.

Table 1b: Oral-laryngeal coordination for different manners of articulation.

		OG-OC			P-OC		
Place	Sess	stop	fricative	P	stop	fricative	P
labial	1	5 (7)	-9 (4)	.000	125 (10)	104 (7)	.000
	2	7 (5)	-14 (8)	.000	111 (7)	95 (9)	.000
alveolar	1	5 (3)	-14 (9)	.000	124 (7)	111 (15)	.002
	2	16 (10)	-6 (6)	.000	115 (7)	105 (13)	.005

In summary, Table 1a,b shows the tendency that the time between consonant onset and glottal opening onset as well as between consonant onset and peak glottal opening is in the order velar > alveolar > labial across stops and fricatives (except that the voiceless velar fricative was not investigated) and in the order stops > fricatives across places of articulation. In a schematical format these general relations are illustrated in Figure 4. Figure 4 shows the supralaryngeal configuration (called oral gesture) of an (in this case word-initial) intervocalic obstruent, with supralaryngeal opening for the first vowel, closure (or constriction in the case of fricatives) for the obstruent, and opening again for the second vowel. Above the oral gesture we find a glottal opening gesture that is coordinated relatively late with respect to the oral gesture, and below we find a glottal gesture with relatively early timing. Although Figure 4 does not represent the facts in detail, it suffices to illustrate the difference between the late timing of the glottal gesture in back places of articulation as opposed to front ones and the late timing of stops as opposed to fricatives (for reasons of space only the P-OC parameter is included in Figure 4; it also turned out to be slightly more robust statistically than OG-OC). For the issue at hand, i.e. the gestural organization

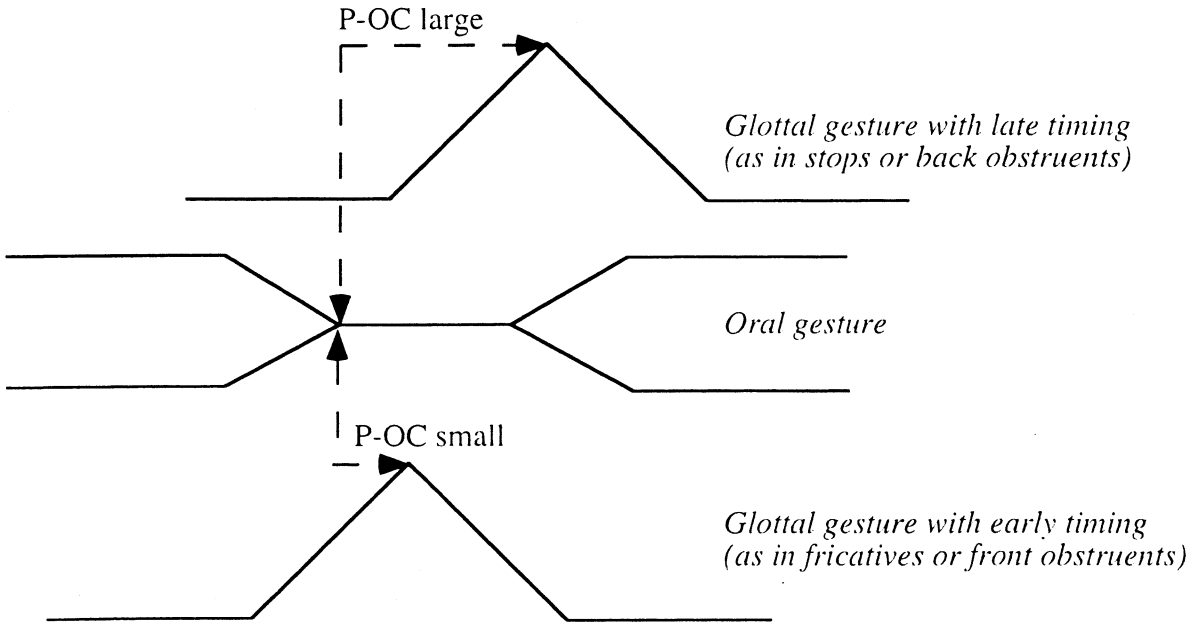


Figure 4: Different patterns of oral-laryngeal coordination.

of obstruent sequences across a word boundary, the different coordination patterns found in Table 1 and illustrated schematically in Figure 4 have the following consequence. Assuming that the sound that precedes the obstruent in question is not a vowel, as in Table 1 and Figure 4, but a voiceless fricative like /ʃ/ (as it is the case in the present investigation) and assuming the constancy

of the glottal opening gesture of /ʃ/ across the specifics of the following sound, two different glottal opening patterns are expected, depending on whether the obstruent after /ʃ/ has a relatively early or late timing of glottal opening (Figure 5). Figure 5 shows two hypothetical glottal opening patterns.

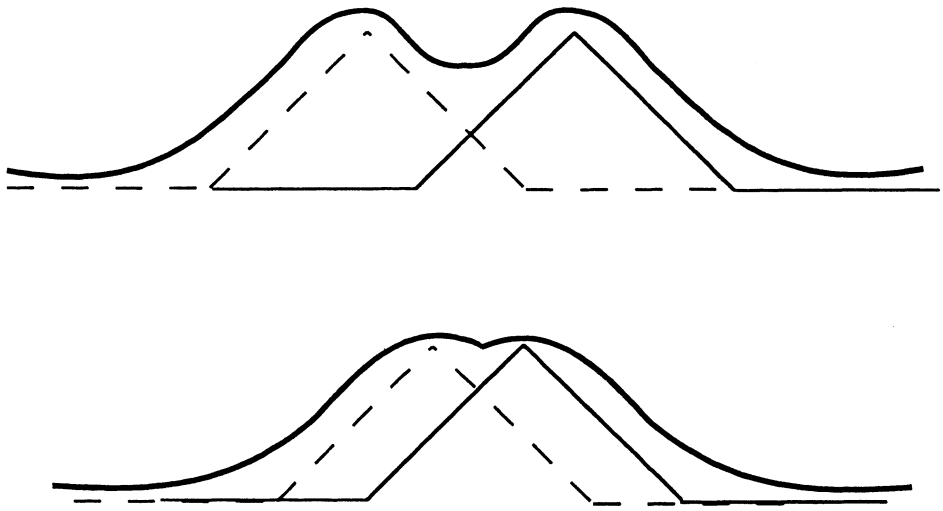


Figure 5: Glottal opening patterns with different timing of the second obstruent.

In the illustration above, the underlying glottal opening of a voiceless fricative like /ʃ/ is combined with the underlying glottal opening of a voiceless obstruent that has a late glottal opening pattern, as in stops and obstruents backwards in the oral cavity. If the two glottal gestures are combined a clear bimodal gesture is expected. In the illustration below, the gesture of the fricative is combined with an early-timed glottal opening found in fricatives or obstruents produced forward in the oral cavity. After combination of the underlying gestures a result is expected in which the bimodal structure is greatly reduced or no longer present. As mentioned earlier in this section and illustrated with Figure 3, the expectations expressed in Figure 5 are borne out as a tendency in the "eye-balling" method pursued here.[4] One possibility of approaching a more quantitative index of gestural aggregation is the use of a velocity calculation of the transillumination signal. If the compound glottal gestures show a clear bimodal pattern two additional zero crossings should occur in the velocity curve (cf. Munhall & Löfqvist 1992 for this criterion). Figure 6 provides an illustration of this procedure. Figure 6 shows the smoothed transillumination signal (above) time-aligned with a velocity calculation (below) of an example of the utterance *rasch Kir*, with the sequence /ʃ#k/. The first zero crossing in the velocity curve is aligned with the first peak, which belongs to /ʃ/. The second zero crossing is aligned with the valley in-between the peaks, and the third one is aligned with the peak that is associated with /k/. Instead of the three zero crossings only one occurs if no two clear peaks exist. Going through the material it also turned out that in a number of cases in which two peaks were identifiable still only a single zero crossing appeared. This was found in examples were the intensity of one of the peaks was considerably lower than that of the other, in which case no direction change occurs in the transillumination signal and subsequently no additional zero crossing in the velocity curve. The number of productions with additional zero crossings found in the corpus is listed in Table 2. Table 2 shows the number of voiceless obstruents occurring after /ʃ/ across a word boundary that were produced with an additional pair of zero crossings in the velocity curve. It is subdivided according to the obstruents involved (columns) and the two different recording sessions (rows). The total number of tokens for each slot is approximately twenty. Thus, for each obstruent and both recordings cases with additional zero crossings turned out to be in the minority. Partially this means that the presence of

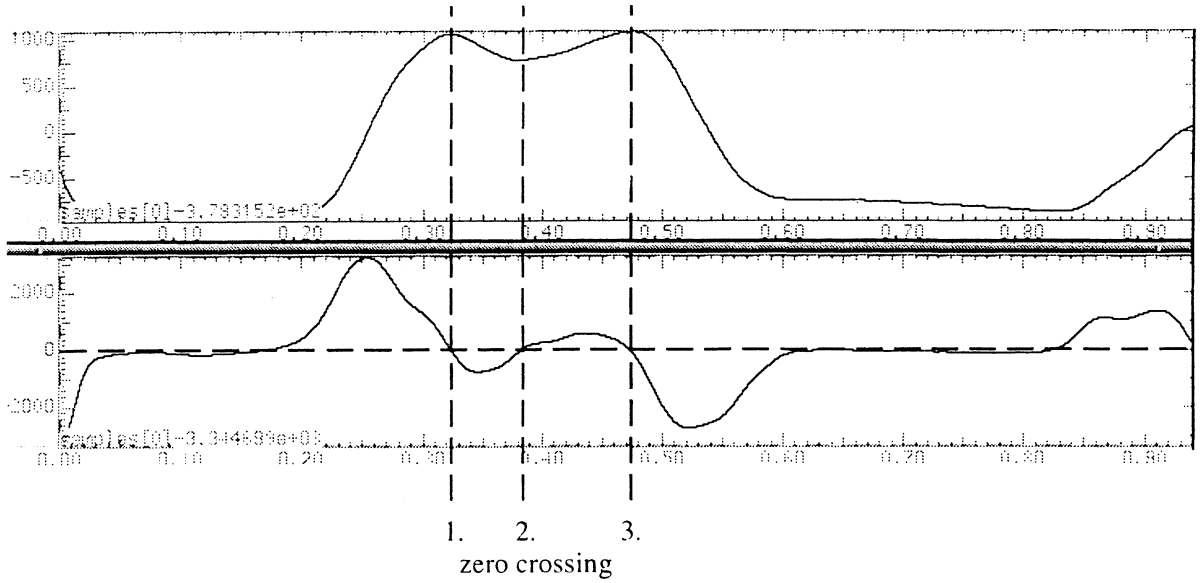


Figure 6: Additional zero crossings in bimodal glottal openings.

additional zero crossings, as objective and methodologically simple it is, might be a criterion that is too conservative for the type of investigation that is pursued here. But this result also indicates that the level of gestural aggregation in this study is relatively high despite the fact that the material was

Table 2: Number of tokens produced with additional zero crossings.

	f	s	p	t	k
Session 1	4	5	5	6	6
Session 2	0	0	1	1	3

spoken in a relatively slow speech rate and nonspontaneous "lab speech" style (cf. Pétursson 1977 for variation of gestural aggregation in lab speech and Fukui & Hirose 1983 for differences between speakers). Thus, the rate of misses, i.e. cases where despite an intervening word boundary no clear two glottal opening peaks can be observed, is relatively high. But despite an overall small number of additional zero crossings, the absolute values shown in Table 2 are actually consistent with the patterns that are expected from the different configurations of oral-laryngeal timing illustrated in Figures 3 and 5. As a tendency, higher numbers of tokens with additional zero crossings are found in stops as compared to fricatives and in consonants produced further back as compared to consonants further to the front of the oral cavity.

6. Further aspects of laryngeal word boundary marking

In this contribution the phonetics of word boundary marking was investigated with respect to a quite specific topic, namely the number of glottal opening peaks in sequences of voiceless obstruents. Even if we restrict ourselves to the role of the larynx in the production of cues to word segmentation there are a number of other potential topics that deserve further attention. However, only a few remarks along these lines can be made here.

Glottal opening behavior has been addressed in this study from a largely qualitative and categorical point of view, when we were looking at the number of glottal opening peaks occurring in obstruent sequences with and without word boundaries. But it is also possible to look for cues to word boundary marking in the *quantitative* and *gradient* properties of glottal opening gestures.

An interesting design for a transillumination study that proceeds along these lines is found in the work of Cooper (1991) for English. By distinguishing between word-initial stressed, word-medial stressed, word-initial unstressed, and word-medial unstressed syllables in his stimuli, Cooper was able to differentiate the effects of word stress from the effects of word segmentation on glottal opening. The variables that were measured by Cooper include the dimensions of the glottal opening gesture alone, such as glottal opening duration and peak glottal opening, and patterns of oral-laryngeal coordination similar to the ones discussed in connection with Table 1.

As another area of laryngeal word boundary marking we may want to look beyond obstruent production and consider the glottal opening behavior of the sound /h/. One aspect that makes /h/ interesting for the topic of word boundary marking is its phonotactic status. The position in which /h/ is found most commonly in Modern Standard German is at the beginning of the word (e.g. *Hafen* 'harbor', *halten* 'hold', *Heirat* 'marriage'). In this position it occurs in isolation before the vowel and may not combine with other consonants. The restriction against a combination of /h/ with other consonants holds for any position within a (monomorphemic) word. The only position in which /h/ can be found word-internally in German is between two vowels, provided that the preceding vowel is tense (or a diphthong) and the following vowel is not [ə] or [ɐ]. The number of words with word-medial /h/ is quite small and many of them are loans (e.g. *Ahorn* 'maple', *Alkohol* 'alcohol', *Uhu* 'eagle owl', *Vehikel* 'vehicle'). It has been pointed out already by Trubetzkoy (1939: 247) that /h/ is an effective boundary marker in German. If a sequence of consonant and /h/ occurs, it can be inferred that a word or morpheme boundary falls in-between the two sounds.[5]

On the phonetic side it is interesting to notice that /h/ differs from voiceless fricatives and aspirated stops by the fact that glottal opening can coincide with continuing vibrations of the vocal folds. In obstruents glottal opening acts as a devoicing mechanism, but in /h/, where no supra-glottal obstruction occurs, voicing can carry through part or all of its glottal opening. This can be seen in the illustrations in Sawashima & Hirose (1983: 17) for Japanese and Löfqvist & McGowan (1992: 98) for a Swedish speaker. Figure 7 shows the same for German (see also Hartmann 1963

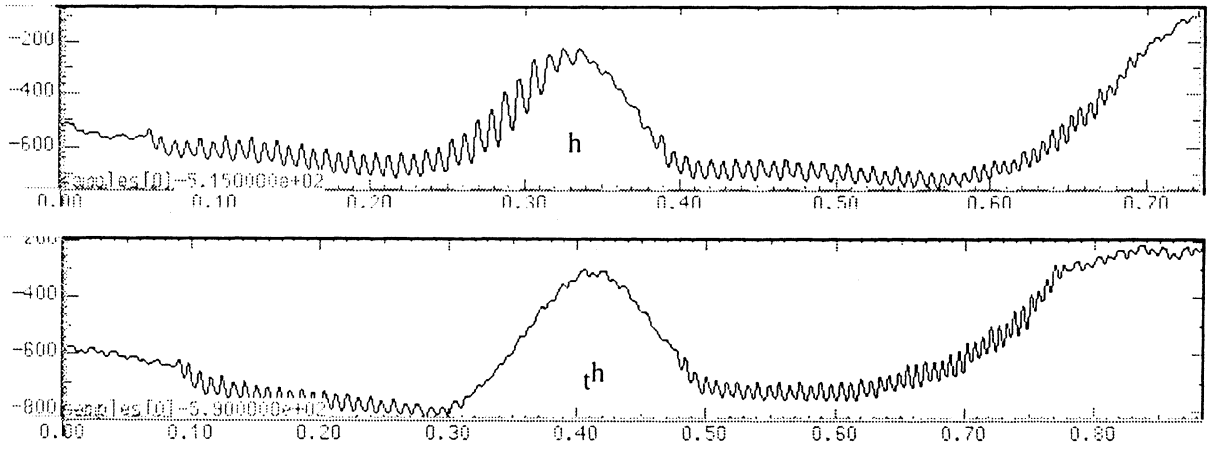


Figure 7: Glottal opening gestures in [h] and [t^h].

and Stock 1971: 100ff. for German). Figure 7 shows the transillumination signal of /h/ in the utterance *nie hier* 'never here' (above) in comparison with the transillumination signal of aspirated /t/ in the utterance *nie Tier* 'never animal' (below). Contrary to the other transillumination examples presented so far, no smoothing was applied to these cases, so that voicing information was preserved. The glottal opening associated with the aspirated stop is almost entirely voiceless, whereas more than half of the glottal opening of /h/ is voiced. In the corpus of /h/ productions there were also tokens with voicing throughout and tokens with less voicing than shown here, but

always with more than in aspirated stops. These fact show that glottal opening gestures with substantial voicing are always indicative of the segment /h/ in German and subsequently, for the phonotactic reasons mentioned above, indicative of word-initial position.[6]

Finally, is should be discussed whether the opposite of glottal opening, namely *glottal constriction*, has potential as an index of word segmentation. One case that has been interpreted frequently as a boundary marker in German phonetics and phonology is the glottal stop (Trubetzkoy 1939: 244f., Moulton 1947, among others). The glottal stop has a distribution in German that is not much different from the occurrence of /h/, discussed above, except that it is probably more sensitive to the degree of stress in the syllable that it initiates (see Moulton 1947, Kohler 1995: 100ff., 168f., Wiese 1996: 58ff.). According to the pronouncing dictionary *Großes Wörterbuch der deutschen Aussprache* (Krech et al. 1982) the glottal stop can occur not only word-initially, but also word-internally between two vowels, if the second vowel is stressed (e.g. *The[ʔ]ater* 'theater', *Di[ʔ]ode* 'diode'), though other pronouncing dictionaries do not agree. Most of these items are loans, analogously to the situation for /h/, though with more examples, provided the presence of glottal stop in this context is in fact a stable phenomenon (which is doubtful in light of the fact that Kohler 1994 found no glottal stops in cases of this type). Contrary to /h/, which is commonly classified as phonemic in German, the glottal stop has a wider range of realizations and more variability in its occurrence than /h/. This is one of the major reasons why the glottal stop is usually not considered phonemic in German. The occurrence and realization of what is understood and transcribed as the glottal stop in German has been investigated in detail by Krech (1968) and Kohler (1994). As one of their results, both found that glottal stops are produced more frequently after voiceless obstruents than after vowels (and sonorants). Notice that, like /h/, the glottal stop does not usually combine with preceding obstruents word-internally (Kohler 1995: 101). This implies that post-voiceless position, which turned out to be particularly favorable for the expression of the glottal stop, is a context in which the glottal stop occurs word-initially (e.g. in *das [ʔ]Ohr* 'the ear').[7]

Glottal stop and the laryngealization that accompanies or replaces it is visible in the transillumination signal as a gesture-like lowering of intensity beyond the baseline that is associated with the adjacent vowels. An illustration of the glottal stop in word-initial position surrounded by vowels is provided in Figure 8 (cf. a similar view of the glottal stop in Löfqvist & McGowan 1992: 98 for a Swedish speaker). Figure 8 shows the audio signal (above) and unsmoothed transillumination signal (below) of a glottal stop found in the utterance *nie ihr* 'never you PL', representative of several repetitions. It can be inferred from Figure 8 that the glottal constriction produced in

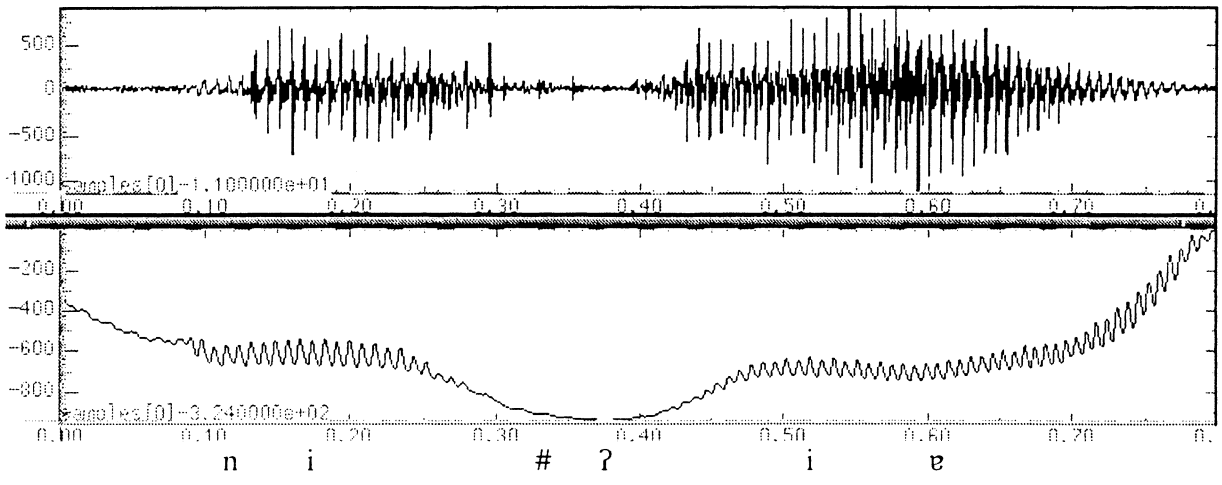


Figure 8: Glottal constriction in [ʔ].

a glottal stop leads to a compression of the vocal fold tissues that reduces the conduction of light through the glottis beyond the level found in the closed portion of a glottal cycle in normal voice production (see the voicing information of the adjacent vowels in the unsmoothed transillumination signal). Given what was said about the distribution of the glottal stop, the occurrence of a glottal constriction pattern beyond the vowel baseline is a likely indication of a word boundary (at least in terms of the notion "phonological word" mentioned in Notes 5 and 7).

So far glottal stop was discussed with respect to word-initial position. Kohler (1994) shows that a glottal stop or related forms of glottal constriction can also occur word-finally, in which case they accompany or even replace stop consonants. Kohler, however, also mentions that this glottal constriction pattern can extend to positions within a word, especially in combination with nasals, like in *Leutnant* 'lieutenant', *hinten* 'behind', *Punkten* 'point DAT PL' (Kohler 1994: 45; boldface indicating the stop that undergoes glottal replacement). In some cases the mentioned notion of "phonological word" might be of help, but not in all, which creates a false alarm problem (i.e. glottal stop, but no word boundary).

7. Conclusion

In this contribution we have been concerned with the question of whether information about the presence or absence of word boundaries is encoded in the way glottal opening gestures are organized in the production of voiceless obstruent sequences. An appealing argument in favor of the "Word Boundary Marking Hypothesis" was offered by the fact that the same sequence of voiceless obstruents of the type /st/ is produced with a single (monomodal) glottal opening word-internally, but with a double (bimodal) glottal opening if the two sounds are separated by a word boundary. From this finding it seemed to follow that the difference between a bimodal and a monomodal glottal opening pattern can be directly attributed to the presence versus the absence of a word boundary, respectively. The flaw in this argument lies in the fact that these two obstruent sequences are not really identical on the phonetic level: if a voiceless stop occurs alone in word-initial position, it is aspirated, but if it clusters with a preceding fricative it is unaspirated (at least this is the regular case in the Germanic languages). With this knowledge in mind it became possible to attribute the different glottal opening patterns to the particular phonetic properties of the segments involved, rather than to an autonomous influence of juncture. In this example, the presence of aspiration could be made responsible for one of the two peaks in the glottal opening pattern. This is the essence of what was referred to as "Löfqvist's rule" in § 2. Löfqvist's rule offered an alternative to the Word Boundary Marking Hypothesis, that did not only offer a different perspective to the interpretation of the /st/ case, but that could also account for cases that constitute a direct falsification of the Word Boundary Marking Hypothesis. Some of these cases were already known in the literature, while others were added by the new evidence on German that was presented here.

If we focus on the role of aspiration in the light of the information discussed and presented here, the most straightforward reasoning seems to be that it is aspiration in itself that carries information about word boundary marking, and that aspiration in turn requires a certain arrangement of glottal opening gestures. In other words, word boundaries are not directly reflected in the organization of glottal gestures, but only indirectly – mediated by aspiration as the most direct "demarcative feature" in this case (to use the terminology of Trubetzkoy 1939 or Jakobson & Waugh 1987). That the presence or degree of aspiration is an important boundary marker in German is for example claimed by the pronouncing dictionary *Duden* (Mangold 1990: 49). Duden claims /p,t,k/ to be strongly or even very strongly aspirated in word-initial position, even if the first syllable in the word is not stressed (as in the initial [t^{h}] of *Talént* 'talent'). Aspiration has been interpreted as juncture-dependent already by Moulton (1947).[8] Aspiration can also be an explanation for glottal opening behavior in other functions than word boundary marking. In § 5 we saw that the clarity with which a bimodal glottal opening pattern occurs is in part determined by the

place of articulation of the following stop (being in the order k > t > p). Yet, still a large amount of free variation was observed. To the extent that glottal organization is interpreted as a "redundant feature" in the expression of place of articulation (cf. Jakobson & Waugh 1987 for this term), it must be acknowledged that it is again aspiration that does the better job as a redundant feature of place of articulation. In the acoustic signal of the transillumination study of Jessen (1995) the order k > t > p in aspiration duration turned out to be a reliable and statistically significant feature of place of articulation in word-initial position, which is a more stable effect than the degree of gestural aggregation, addressed in § 5.

Thus, one way we can conclude this paper is by hypothesizing that in the expression of word boundaries (as well as in some other functions) the *acoustics* (here: aspiration) has primacy over the *articulation* (here: organization of glottal opening gestures). Such a conclusion would emphasize the goal-oriented or listener-oriented nature of articulatory (here: gestural) organization in general (cf., among others, Jakobson & Waugh 1987, Lindblom 1990, Kohler 1994, Perkell et al. 1995).

8. References

Browman, C.P. & Goldstein, L. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3: 219-252.

Browman, C.P. & Goldstein, L. 1992. Articulatory phonology: an overview. *Phonetica* 49: 155-180.

Cooper, A.M. 1991. *An articulatory account of aspiration in English*. Ph.D. Dissertation, Yale University.

Davis, K. 1994. Stop voicing in Hindi. *Journal of Phonetics* 22: 177-193.

Fischer-Jørgensen, E. & Hutters, B. 1981. Aspirated stop consonants before low vowels, a problem of delimitation - its causes and consequences. *Annual Report of the Institute of Phonetics of the University of Copenhagen* 15: 77-102.

Frøkjaer-Jensen, B., Ludvigsen, C. & Rischel, J. 1971. A glottographic study of some Danish consonants. In Hammerich, L.L., Jakobson, R. & Zwirner, E. (eds.) *Form and substance. Phonetic and linguistic papers presented to Eli Fischer-Jørgensen*. Odense: Akademisk Forlag. 123-140.

Fukui, N. & Hirose, H. 1983. Laryngeal adjustments in Danish voiceless obstruent production. *Annual Bulletin. Research Institute of Logopedics and Phoniatrics. University of Tokyo* 17: 61-71.

Giegerich, H.J. 1989. *Syllable structure and lexical derivation in German*. Bloomington: Indiana University Linguistics Club.

Hall, T.A. 1992. *Syllable structure and syllable-related processes in German*. Tübingen: Niemeyer.

Halle, M. & Stevens, K.N. 1971. A note on laryngeal features. *MIT Research Laboratory of Electronics Quarterly Status Report* 101: 198-213.

Hartmann, E. 1963. Positionsbedingte phonetische Varianten des deutschen Hauchlautes. *Zeitschrift für Phonetik, Sprachwissenschaft und Komunikationsforschung* 16: 49-55.

Hayes, B. 1986. Inalterability in CV Phonology. *Language* 62: 321-351.

Hoole, P., to appear. Laryngeal coarticulation: coarticulatory investigations of the devoicing gesture. In Hardcastle, W.H. & Hewlett, N. (eds.) *Instrumental studies of coarticulation*. Cambridge etc.: Cambridge University Press.

Hoole, P., Pompino-Marschall, B. & Dames, M. 1984. Glottal timing in German voiceless obstruents. *Proceedings of the International Congress of Phonetic Sciences* 10, 2b: 399-403.

Iverson, G.K. & Salmons, J.C. 1995. Aspiration and laryngeal representation in Germanic. *Phonology* 12: 369-396.

Jakobson, R. & Waugh, L.R. 1987 (2). *The sound shape of language*. Berlin etc.: Mouton de Gruyter.
Jessen, M. 1995. Glottal opening in German obstruents. *Proceedings of the International Congress of Phonetic Sciences* 13, 3: 428-431.
Jessen, M., to appear. *Phonetics and phonology of tense and lax obstruents in German*. Amsterdam: Benjamins.
Kenstowicz, M. 1994. *Phonology in generative grammar*. Cambridge, USA & Oxford, UK: Blackwell.
Kim, C.-W. 1970. A theory of aspiration. *Phonetica* 21: 107-116.
Klatt, D.H. 1976. Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America* 59: 1208-1221.
Knetschke, E. & Sperlbaum, M. 1987. *Zur Orthoepie der Plosiva in der deutschen Hochsprache. Eine auditiv-komparative Untersuchung*. Tübingen: Niemeyer.
Kohler, K.J. 1994. Glottal stops and glottalization in German. *Phonetica* 51: 38-51.
Kohler, K.J. 1995 (2). *Einführung in die Phonetik des Deutschen*. Berlin: Erich Schmidt Verlag.
Krech, E.-M. 1968. *Sprechwissenschaftlich-phonetische Untersuchungen zum Gebrauch des Glottisschlageinsatzes in der allgemeinen deutschen Hochlautung*. Basel & New York: S. Karger.
Krech, E.-M. et al. 1982. *Großes Wörterbuch der deutschen Aussprache*. Leipzig: VEB Bibliographisches Institut.
Ladefoged, P., Williamson, K., Elugbe, B. & Sister A.A. Uwalaka. 1976. The stops of Owerri Igbo. *Studies in African Linguistics* Supplement 6: 147-163.
Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W.J. & Marchal, A. (eds.) *Speech production and speech modelling*. Dordrecht: Kluwer. 403-439.
Lindqvist, J. 1972. Laryngeal articulation studied on Swedish subjects. *Speech Transmission Laboratory Charterly Progress and Status Report. Royal Institute of Technology, Stockholm* 2-2: 10-27.
Lisker, L. & Baer, T. 1984. Laryngeal management at utterance-internal word boundary in American English. *Language and Speech* 27: 163-171.
Löfqvist, A. 1978. Laryngeal articulation and junctures in the production of Swedish obstruent sequences. In Gårding, E., Bruce, G. & Bannert, R. (eds.) *Nordic Prosody*. Lund: Lund University Press. 73-83.
Löfqvist, A. 1992. Acoustic and aerodynamic effects of interarticulator timing in voiceless consonants. *Language and Speech* 35: 15-28.
Löfqvist, A. & Yoshioka, H. 1980. Laryngeal activity in Swedish obstruent clusters. *Journal of the Acoustical Society of America* 68: 792-801.
Löfqvist, A. & Yoshioka, H. 1981. Laryngeal activity in Icelandic obstruent production. *Nordic Journal of Linguistics* 4: 1-18.
Löfqvist, A. & McGowan, R.S. 1992. Influence of consonantal environment on voice source aerodynamics. *Journal of Phonetics* 20: 93-110.
Lotzmann, G. 1975. *Zur Aspiration der Explosivae im Deutschen. Ein sprechwissenschaftlich-phonetischer Beitrag zur deutschen Hochlautung*. Göppingen: Kümmerle.
Mangold, M. 1990 (3). *DUDEN Aussprachewörterbuch*. Mannhein etc.: Dudenverlag.
McCarthy, J.J. 1988. Feature geometry and dependency: a review. *Phonetica* 43: 84-108.
Moulton, W.G. 1947. Juncture in Modern Standard German. *Language* 23: 212-226.
Munhall, K. & Löfqvist, A. 1992. Gestural aggregation in speech: laryngeal gestures. *Journal of Phonetics* 20: 111-126.
Nespor, M. & Vogel, I. 1986. *Prosodic phonology*. Dordrecht: Foris.

Perkell, J.S., Matthies, M.L., Svirsky, M.A. & Jordan, M.I. 1995. Goal-based speech motor control: a theoretical framework and some preliminary data. *Journal of Phonetics* 23: 23-35.

Pétursson, M. 1977. Timing of glottal events in the production of aspiration after [s]. *Journal of Phonetics* 5: 205-212.

Sawashima, M. & Hirose, H. 1983. Laryngeal gestures in speech production. In MacNeilage, P.F. (ed.) *The production of speech.* New York etc.: Springer. 11-38.

Stock, D. 1971. *Untersuchungen zur Stimmhaftigkeit hochdeutscher Phonemrealisationen.* Hamburg: Buske.

Trubetzkoy, N.S. 1939. *Grundzüge der Phonologie.* 1958 by Göttingen: Vandenhoeck & Ruprecht.

Wiese, R. 1996. *The phonology of German.* Oxford: Oxford University Press.

Yoshioka, H., Löfqvist, A. & Hirose, H. 1981. Laryngeal adjustments in the production of consonant clusters and geminates in American English. *Journal of the Acoustical Society of America* 70: 1615-1623.

Yoshioka, H., Löfqvist, A. & Hirose, H. 1982. Laryngeal adjustments in Japanese voiceless sound production. *Journal of Phonetics* 10: 1-10.

Yoshioka, H., Löfqvist, A. & Collier, R. 1982. Laryngeal adjustments in Dutch voiceless obstruent production. *Annual Bulletin. Research Institute of Logopedics and Phoniatrics. University of Tokyo* 16: 27-35.

Yu, S.-T. 1992. *Unterspezifikation in der Phonologie des Deutschen.* Tübingen: Niemeyer.

---

* The entire transillumination experiment, of which only certain aspects are presented and discussed here, was supported by Grant DC-00865 from the National Institute on Deafness and Other Communication Disorders to Haskins Laboratories (cf. Jessen 1995, Jessen, to appear).

[1]Sequences of identical adjacent obstruents are refered to as "geminates" in the literature discussed here, even if a word boundary intervenes (see in particular Yoshioka, Löfqvist & Hirose 1981). In most of the phonological literature the term "(true) geminate" is restricted to the case that no morpheme- or word boundary intervenes between the identical obstruents (see Hayes 1986). In this sense the word-internal identical sequences of Japanese, investigated with transillumination and other methods by Yoshioka, Löfqvist & Hirose (1982), constitute true geminates in contrast to the "fake geminates" of English, Swedish etc. However, with respect to the presence of a single glottal opening gesture the Japanese true geminates in most cases behave like the fake geminates reported for the Germanic languages.

[2]The author favors a definition of aspiration duration in which the end of aspiration is not measured as the beginning of voicing in the following vowel (i.e. positive Voice Onset Time), but as the end of aspiration turbulence into the following vowel. One practical approximation of this latter event is the beginning of the second formant in the following vowel (cf. Fischer-Jørgensen & Hutters 1981, Davis 1994, Jessen to appear). Among the disadvantages of the VOT concept is the difficulty of measuring aspiration in languages with a voiced aspirated stop category (Ladefoged et al. 1976, Davis 1994 about this point).

[3]Another timing parameter – P-R (interval between stop release and peak glottal opening) – was statistically more robust than OG-OC and P-OC in the expression of place of articulation. For the word-initial context discussed here P-R was in the order velar > alveolar > labial and turned out to be well correlated with aspiration duration (Jessen 1995, to appear; similarly Cooper 1991 for English). P-R was not listed in Table 1 because it is not applicable to fricatives.

[4]An early coordination of glottal opening in /p/ and the subsequent merger with the glottal opening of a preceding /s/ across a word boundary is also reported by Lisker & Baer (1984) for English.

[5]To say that /h/ is common word-initially is insufficient insofar as /h/ is also common at the beginning of certain prefixes and suffixes (e.g. *hin-*, *her-*; *-haft*, *-heit*), as well as at the beginning of lexical stems, whether occuring in isolation or in combination with other morphemes. One might argue on the basis of these facts that /h/ is more of an index for morpheme boundaries than for word boundaries. However, there is evidence that the appropriate domain is larger than the morpheme and constitutes what is referred to as the "phonological word" by several phonologists.

The occurrence of /h/ is also influenced by the stress level of the syllable it initiates (although there are exceptions to this generalization). For this reason it has been suggested that the "foot" might be more adequate to characterize the domain of the occurrence of /h/ than the phonological word (see Wiese 1996: 60). Althogether, word boundary sensitivity and stress sensitivity are often related in intricate ways in languages such as English or German (Wiese 1996: 72; cf. also Cooper 1991 and Note 7). Notice also that in prosodic phonology the "foot" is not independent of the "phonological word", but strictly dominated by it. This implies that the beginning of the phonological word is also the beginning of a foot (under the application of the "strict layer hypothesis"; see Nespor & Vogel 1986). Thus, both word boundary sensitivity and stress sensitivity are encoded in the concept of the foot in prosodic phonology.

[6]In the example shown in Figure 7 word-initial /h/ occurs between two vowels. Firstly, we need to keep in mind that in words like *Uhu*, *Ahorn*, mentioned above, /h/ also occurs intervocalically and is likely to be produced with substantial voicing. Thus, these words would constitute false alarms to word boundary marking in the sense that in those cases a voiced glottal opening gesture does not indicate a word boundary. Secondly, word-initial /h/ is not voiced to the same extent or not at all when preceded by a voiceless obstruent such as /ʃ/. This is not shown here, but found in productions of the utterance *rasch hier* 'quickly here' (but cf. Stock 1971: 100ff. for several cases of voiced /h/ even in this context).

[7]Again, we need to keep in mind that this is a simplification insofar as the glottal stop can also occur at the beginning of prefixes and stems (though not suffixes) that occur word-internally. Some phonologists generalize the appropriate position as the "phonological word" (see the discussion in Wiese 1996: 72), while others prefer the "foot" as the most appropriate domain for glottal stop insertion, due to its stress sensitivity (cf. Note 5). Proponents of this solution include Giegerich (1989: 62ff.), Hall (1992: 58f.), Yu (1992: 84ff.), and Wiese (1996: 58ff.).

[8]However, the value of aspiration as a word boundary marker in German has to be seen as a gradient, not a categorical phenomenon. That is, aspiration is usually longer word-initially than word-medially, but it is not the case that it is present in the former and absent in the latter context. There is sufficient evidence that a non-negligible amount of aspiration duration exists in word-medial position before schwa (see Jessen, to appear for data and discussion of the literature). It should also be mentioned that along with glottal stop insertion, aspiration has been analyzed as foot-dependent by several phonologists (see the references in Note 7).

From canonical word forms to reduced variants: A study of assimilation and elision in German.

Bernd J. Kröger
Institut für Phonetik der Universität zu Köln

Abstract

In the framework of Articulatory Phonology (Browman and Goldstein 1992) the variety of discrete segmental changes describing the transition from canonical word forms to reduced variants (i.e. elision and assimilation phenomena) can be accounted for by two continuous and non-discrete gestural alteration processes: increase in overlap and decrease in temporal extent of articulatory gestures.
It can be shown that many segmental phenomena like elisions and assimilations in German can be ascribed to these two basic gestural alteration processes. But some assimilation phenomena (progressive and regressive assimilation of place and regressive assimilation of manner) can be described only by introducing a discrete gestural process: gestural (or articulatory) reorganization.
Further we will show that both continuous gestural processes are strongly related to each other. Increase in overlap can be attributed to reduction of temporal extent of gestures if basic gestural association relations are taken into account. In order to develop a comprehensive theory of reduction, we will illustrate that all continuous and discrete gestural processes can be seen as consequences of minimizing articulatory effort.

1 A brief introduction to Articulatory Phonology

1.1 The gesture as a phonetic and phonological unit

Gestures are the basic units of Articulatory Phonology (Browman and Goldstein 1992). They are units of articulatory activity, realizing linguistically relevant vocal tract constrictions like "labial closure" or "glottal opening". Consequently gestures are phonetic as well as phonological units. On one hand gestures are distinctive units and define discrete phonological categories like [place], e.g. labial vs. apical gestures, [manner], e.g. full-closing gestures (for plosives or nasals) vs. near-closing gestures (for fricatives), or [voice], e.g. occurrence vs. no occurrence of a glottal or velic opening gestures. On the other hand each gesture represents a family of functionally equivalent articulatory movement patterns that are actively controlled with reference to speech relevant goals, i.e. the formation of vocal tract constrictions (Saltzman and Munhall 1989). Consequently in the framework of Articulatory Phonology we have no separation of phonological and phonetic units as occurring in segmental theories (e.g. the separation between phoneme and sound).
There are a lot of reasons, which illustrate the importance of the gesture. Firstly - as illustrated above - the gesture is a phonological as well as a phonetic unit. Consequently there is no need to define a phonetic-phonological interface in this approach. The concept "gesture" can be used both in phonetic *and* phonological investigations. In a quantitative model of speech pro-

duction (chapter 2) a phonological description of a gesture can be transformed into an equivalent phonetic one by specifying values for continuous phonetic parameters like target location, strength of gestural activation, and duration of gestural activation (chapter 2.1). Secondly, articulatory measurements indicate that the spatio-temporal structure of articulatory transitions - i.e. the articulatory shape of gestures - is more stable than the spatio-temporal structure in the region of articulatory targets (i.e. around the maxima and minima of articulatory trajectories) (Fujimura 1981 and 1986). Thirdly, a variety of different discrete segmental changes (e.g. elisions and assimilations) occurring in reduced formes (e.g. words in unstressed positions, at high speech rate, or in casual speech) can be ascribed to few continuous gestural processes: increase in overlap of two gestures and decrease of temporal extent of a gesture (Browman and Goldstein 1989 and 1990, Kröger 1993, and this paper, chapter 3). No discrete gestural change - especially no deletion of gestures - occurs in the case of reduction. This is promising since the degree of reduction can result from varying paralinguistic parameters like speech rate. And a variation of a continuous paralinguistic parameter should not lead to a discrete change of linguistic units. Fourthly, the gesture can be seen as a unit of speech production as well as of speech perception. The motor theory of speech perception (Liberman and Mattingly 1985) defines the gesture as its central unit.

1.2 The gestural organisation of a word

In order to understand the gestural approach it is important to understand how an utterance (or at least a word) is organised in the gestural concept. Figure 1 indicates the phonological specification (i.e. the gestural score) of the German word "Kompaß" (compass) in the framework of Articulatory Phonology. Three types of gestures must be differentiated: tract-shaping gestures (TSG), constriction-forming gestures (CFG), and opening gestures (OPG). Gestures can be phonologically specified by four-letter-symbols: {oldl} or {aldl} are dorsal-labial gestures for the realization of the German lax /ɔ/ or lax /a/, {fcla} or {fcdo} are labial or dorsal full-closing gestures, {ncal} is a alveolar near-closing gesture, and {opgl} or {opve} are velic or glottal opening gestures ("velic" is chosen as a term for the active articulator velum while the term "velar" indicates a (passive) place of articulation in our approach). All gestures can be ordered in different gestural tiers as function of their type. Association lines indicate which gesture is timed or phased with respect to which other gesture.

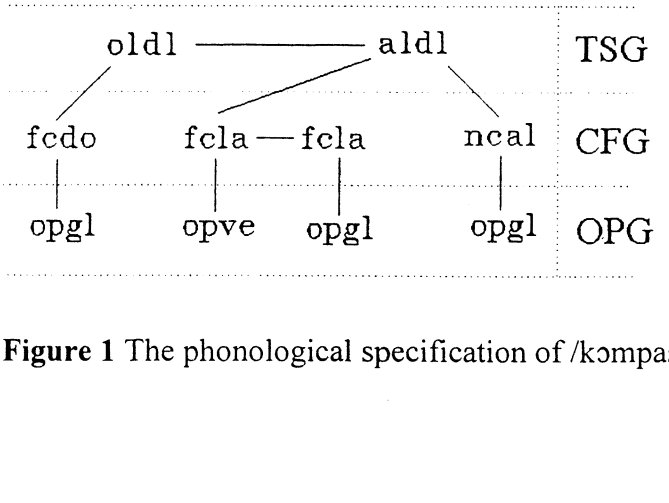


Figure 1 The phonological specification of /kɔmpas/.

2 A gestural speech production model

In the gestural approach any phonological specification can be realized phonetically. A phonetic speech production model has been developed in order to generate the articulation and the acoustic speech signal for a given discrete gestural specification. The first step is the transformation of the phonological gestural specification (four-letter-abbreviations, fig. 1) into a phonetic specification by specifying the values of all (phonetic) parameters for all gestures of an utterance.

2.1 Gestural parameters

Each gesture is defined phonetically (1) by the articulator(s) executing the gestural movement, by target location(s) indicating the gestural target shape(s) or location(s) which is (are) approximated by the gesture-executing articulators(s), (2) by the temporal location and duration of the gestural activation interval, i.e. of the time interval in which the gesture is actively controlling the articulator(s), and (3) by the strength of gestural activation. The gestural parameter "gesture-executing articulator" can be taken directly from the phonological specification of the gesture (chapter 1.2). The parameter "gestural target" is quantitatively defined by specifying values for control parameters like lip protrusion, tongue position, or glottal aperture. These control parameters and its range are model-specific (e.g. Kröger 1993). All other gestural parameters - i.e. the parameter "associated gesture" (e.g. {fcdo} for {oldl} or {opgl} for {fcdo} in "Kompaß", fig. 1) and the three continuous gestural parameters "eigenperiod", "release phase", and "association phase" - specify the strength, temporal location, and temporal extent of the gestural activation interval. Eigenperiod determines the strength of gestural activation; Eigenperiod together with release phase determines the length of gestural activation (Kröger 1993, Kröger et al 1995 and chapter 2.2); Association phase determines the temporal location of the gestural activation interval relative to the location of the associated gesture.

Figure 2 gives the temporal location and the extent of gestural activation intervals for "Kompaß". The gestures are ordered here in five articulatory tiers according to the gesture-performing articulators, i.e. tongue body (TB), tongue tip (TT), lips (LI), velum (VE) and glottis (GL). This figure illustrates two main conditions for an articulator: (1) If gestural activation occurs the articulator is controlled by a gesture. In this case the articulator performs a movement towards the gestural target. (2) If no gestural activation occurs for an articulator this articulator performs a movement towards its inherent neutral position. The neutral position of all articulators defines the production state of a voiced non-nasalized schwa-sound.

Furthermore, this figure shows that gestures overlap in time. Especially different types of gestures overlap considerably in time: Tract forming gestures always overlap with constriction-forming gestures, and constriction-forming gestures always overlap with opening gestures. But also constriction-forming gestures and opening gestures can overlap in time.

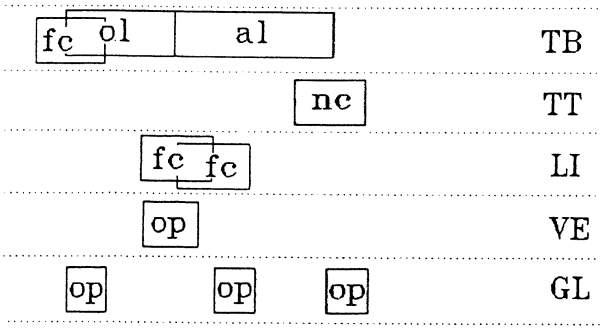


Figure 2 The temporal location and extent of the activation intervals for all gestures of "Kompaß". The abscissa represents time. Each box marks the beginning and ending of a gestural activation interval.

Phonetically the gesture can be seen as a unit of articulatory control. If all gestural parameters are specified, a gesture leads to a defined articulatory movement. Consequently, the specification of all gestures of an utterance leads to an explicit description of its articulation. A vocal tract model can be driven by the gestural specification which generates vocal tract shapes as function of time and subsequently the acoustic speech signal of the utterance. Figure 3 indicates the articulation and the acoustic speech signal for the word "Kompaß" realized in our production model (Kröger 1993). Control parameters and their values are defined with respect to this production model. The control parameters in figure 3 are tongue height (TH), tongue position (TP), tongue tip height (TTH) lip aperture (LA), velic aperture (VA), and glottal aperture (GA).

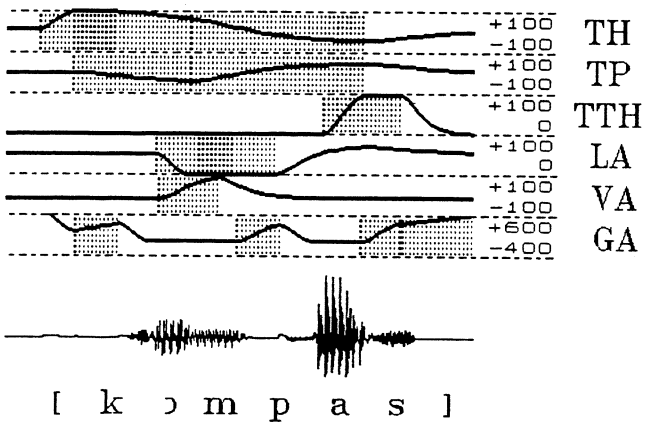


Figure 3 Control parameter time functions (thick lines), gestural activation intervals (shaded areas), and the oscillogram of the synthetic audio signal for "Kompaß". The last glottal opening gesture is followed by a postphonatory opening gesture.

2.2 The dynamic concept for gestures

The dynamics of each gesture determines the spatio-temporal shape of a gesture (i.e. the gestural time function or the gestural movement pattern) and can be generated by a critically damped harmonic oscillator (Saltzman and Munhall 1989, Browman and Goldstein 1990, Kröger 1993, Kröger et al. 1995). In this case the gestural articulator movement is the pattern of an exponential time function asymptotically descending to zero-displacement, i.e. to the gestural target. Examples for gestural movement patterns are shown in figure 4. Here the abscissa indicates relative time values (i.e. phase values, see below) and the ordinate indicates the articulator-target displacement relative to initial displacement. 0% relative articulator-target displacement indicates the target location and 100% indicates initial displacement.
One important parameter of a harmonic oscillator is eigenperiod (i.e. the reciprocal of eigenfrequency) which indicates the level of activation of the oscillator. Low eigenperiod (high eigenfrequency) indicates high activation and vice versa. In the case of critical damping the degree of activation (i.e. eigenperiod) defines the time interval needed for reaching a definite (small) relative articulator-target distance. A gesture with a low eigenperiod value reaches a defined (small) articulator-target distance faster than a gesture with a high eigenperiod value (fig. 4). A relative time scale - the phase scale - can be defined for each gesture if the strength of gestural activation is known: According to the eigenperiod of the gesture the distances on the phase scale are large for low and small for high eigenperiod. In figure 4 the phase scales are indicated for both gestures: above for the dashed lined gesture and below for the solid lined gesture. The figure shows that phase values depend solely on the relative articulator-target-distance: *Phase values indicate the degree to which a gesture is performed.*

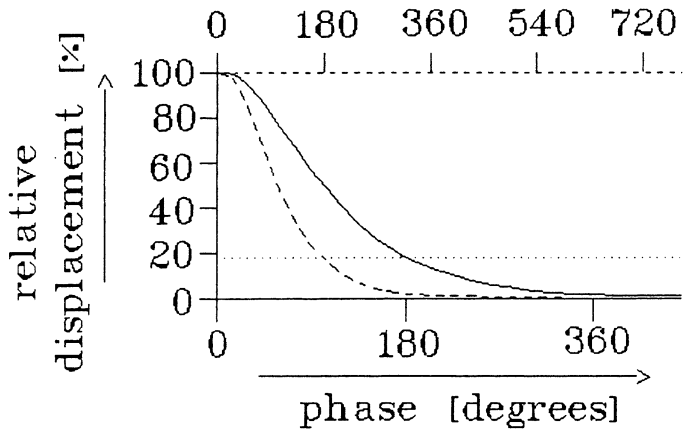


Figure 4 Gestural time functions. The eigenperiod value is lower in the dashed lined than in the solid lined time function. The steady state portion of the gesture occurs below and the transient portion of the gesture above the dotted horizontal line.

As a rule of thumb it can assumed that the gestural target region - i.e. the quasi steady state portion of a gesture - is reached for each gesture at about 180 degrees. For (consonantal) constriction-forming gestures this phase value indicates the beginning of the consonantal obstruction (e.g. full or near closure). Thus the rapid articulator movement towards the gestural target, i.e. the *transient portion* of a gesture, takes place at phase values below 180 degrees

whereas the *quasi steady state portion* in which the articulator is near the gestural target (relative articulator-target distance is lower than approximately 20%, see fig. 4) appears at phase values above 180 degrees.
The remaining gestural parameters are defined with respect to the gestural phase scale. The release phase value indicates the final degree of realization of a gesture, i.e. the degree of articulatory undershoot of a gesture. If for example the release phase value of a consonantal gesture is lower than 180 degrees no consonantal closure will be produced. And the longer a gesture is activated above 180 degrees the more quasi steady state portion of the gesture (e.g. the more time portion of a consonantal closure, of a quasi steady (vocalic) vocal tract shape, or of a glottal or velic opening) is produced. Together with eigenperiod the release phase value determines the temporal extent of a gestural activation interval.
The association phase value of a gesture determines the position of the phase scale of this gesture with respect to the time instant defined by the association line (fig. 1). If for instance the association phase value is 0 degrees, the gesture starts at the time instant defined by the association line; if the association phase value is 180 degrees for a (consonantal) constriction-forming gesture, this gesture is timed (or phased) with respect to the beginning of its consonantal obstruction.

2.3 Gestural phasing rules

The association phase value determines the temporal location of a gesture (i.e. the position of its phase scale) with respect to the time instant defined by the association line. In order to determine gestural phasing completely, association rules are added defining which gesture has to be phased with respect to which other gesture (Browman and Goldstein 1990). Consequently, these rules define the time instants of phasing, i.e. the time instants represented by association lines. Four association rules can be established: (1) Each vocalic gesture is phased with respect to the offset of the preceding vocalic gesture (horizontal phasing line in the tract-shaping tier in fig. 1). (2) The first consonantal gesture of a consonant cluster is phased with respect to the onset of the syllable-defining vocalic gesture if the cluster is syllable-initial, and with respect to its offset if the cluster is syllable-final (transversal phasing lines from the tract-shaping to the constriction-forming tier in fig. 1). (3) Non-first consonantal gestures of a consonant cluster are phased with respect to the offset of the preceding consonantal gesture within this cluster (horizontal phasing line in the constriction-forming tier in fig. 1). (4) Opening gestures are phased with respect to the offset of the pertinent consonantal gesture (vertical phasing lines from the constriction-forming tier to the tier of opening gestures in fig. 1).
The association phase values of each gesture, together with these association rules, determine the intergestural constellation completely. The rules given above together with the specification of association phase values lead to four main principles for gestural coordination: (1) Vocalic gestures are in an immediate succesion without gaps (rule 1 and association phase value of zero). They act as a "ground" to consonantal "figures" (Browman 1991). The articulatory movements resulting from these series of vocalic gestures are comparable to the "vocalic base function" in Fujimura's C/D model (Fujimura 1992) or to the "vowel component" in Öhman's model (Öhman 1967). Consequently, consonantal gestures are completely overlapped by vocalic gestures. (2) The consonantal obstruction interval of a consonant (cluster) coincides with the transient portion of a vowel gesture (rule 2 and association phase value of 180 degrees). Thus the vocalic transition portions of the tongue body are hidden by consonantal constric-

tions. (3) Consonantal obstructions within a consonant cluster are produced without gaps (rule 3 and association phase value of 180 degrees) as such gaps would be perceived as interconsonantal vocalic segments. (4) The temporal extent of the activation interval of an opening gesture coincides with the consonantal obstruction interval (rule 4 and association phase value of 180 degrees). This rule ensures the proper "intrasegmental" timing in the case of voiceless sounds or nasals. In our elaborated quantitative model this rule is differentiated for plosives, fricatives, and nasals according to the articulatory measurements of Löfqvist and Yoshioka (1984):

2.4. Comparison of segmental and gestural approach

Firstly the gestural approach can be interpreted as a non-segmental concept. Gestures are the central units of phonological as well as phonetic description. While segments are serially ordered in segmental approaches gestures are ordered on parallel gestural or articulatory tiers (fig. 1 and fig. 2). Furthermore it is an important feature of this approach that gestures overlap in time. This overlap of gestures in time and the non-serial ordering of gestures is an important feature of the gestural approach. Consequently Articulatory Phonology belongs to non-linear and non-segmental phonologies.
Secondly, it should be mentioned that the concept of "coarticulation" is defined in different ways in segmental theories. The gestural approach allows the replacement of the concept "coarticulation" by the concept of "gestural coproduction". Temporal overlap is clearly defined in this qualitative and quantitative concept and consequently gestural overlap can be concretely measured in this approach.
Thirdly, it is an advantage of the gestural approach that a discrete phonological specification can be transformed into a concrete phonetic realization by means of our gestural production model as is illustrated here for the example "Kompaß" (fig. 1, 2, and 3). Articulator movements and the acoustic signal can be generated from a phonological specification in Articulatory Phonology. This provides us the possibility of perceptual evaluation of gestural specifications and also of gestural processes as has been taken advantage of in this study (see chapter 3).

3 Assimilations and elisions in the gestural framework

3.1 Assimilations and elisions in German and gestural processes

As a consequence of reduction in connected speech a lot of segmental changes - mainly assimilations and elisions - can be found. The main hypothesis of Articulatory Phonology is that these different kinds of *discrete segmental changes* can be ascribed to few simple *continuous gestural alteration processes*, i.e. increase in temporal overlap of gestures and decrease of the temporal extent of a gesture (Browman and Goldstein 1990). It is important to emphasize that these discrete segmental changes can be realized without deletions of any gesture.
In our investigation we verified this hypothesis in the case of assimilations (e.g. assimilation of place, manner, nasality, or voice) and elisions (e.g. elision of [ə]or of [t]) occurring in German (Kohler 1995, p. 205ff).

The cases investigated by using our gestural production model are listed in table 1. It must be emphasised, that in some cases the forms on the left side of table 1 (starting forms) are reduced forms (e.g. in the case of reduction of double consonants in "kommen" [mm]->[m]: the reduction of double consonants is preceded by elision of [ə] and progressive assimilation of nasality. So the starting form is [kɔmm̩]).

Category of segmental change	examples investigated by using the gestural production model
Elision of [ə]	"eben" [bən]->[bn], "reden" [dən]->[dn], "legen" [gən]->[gn], "Adel" [dəl]->[dl]
Elision of [t]	"Glanz" [nts]->[ns], "erhältst" [lts]->[ls], "restlich" [stl]->[sl], "rechtlich" [çtl]->[çl]
Reduction of double consonants	"mitteilen" [tt]->[t], "komm(e)n" [mm]->[m], "wegkommen" [kk]->[k]
Progressive assimilation of place	"eb(e)n" [bn]->[bm], "komm(e)n" [mn]->[mm], "Beamt(e)n" [mtn]->[mpm], "verlog(e)n" [gn]->[gŋ]
Regressive assimilation of place	"mit mein(e)m" [nm]->[mm], "mit jed(e)m" [dm]->[bm], "mit fett(e)m" [tm]->[pm]
Regressive assimilation of manner	"das Schiff" [sʃ]->[ʃʃ], "Eisschrank" [sʃ]->[ʃʃ]
Progressive assimilation of nasality	"umbenennen" [mb]->[mm], "Bundes" [nd]->[nn], "angegeben" [ŋg]->[ŋŋ]
Regressive assimilation of nasality	"Agnes" [gŋ]->[ŋŋ], "eb(e)n" [bm]->[mm], "wird(e)n" [dn]->[nn]
Progressive assimilation of voicelessness	"ratsam"[tz]->[ts], "dasselbe" [sz]->[ss], "das Bad" [sb]->[sb̥], "wegbringen" [kb]->[kb̥]
Sonorization (i.e. assimilation of voice)	"muß ich" [s]->[z], "hat er" [t^h]->[d]
Reduction of degree of opening	"ich habe" [b] -> [β_T], "ich lege" [g]->[γ_T]

Table 2 Categories of segmental changes and concrete examples for each category investigated by using our gestural speech production model.

During the procedure of generation of the segmental changes we firstly generated the unreduced form (starting form) of these words. Secondly we tried to generate the reduced forms by identifying and applying the appertaining gestural alteration process for each word. Thirdly transcriptions of both acoustic forms were analysed to find out whether the segmental change has occurred.

We identified underlying gestural alteration processes in the case of 8 of the given 11 categories of segmental changes, i.e. (1) in the case of elision of [ə] and (2) of [t], (3) in the case of reduction of double consonants, (4) in the case of regressive and (5) progressive assimilation of nasality, (6) in the case of progressive assimilation of voicelessness, (7) in the case of sonorization and (8) in the case of reduction of degree of opening. The appertaining gestural

alteration processes are indicated in figure 5 for one example for each category of segmental change. The figure shows the gestural score before (left side) and after (right side) each segmental change. The altered gestures are indicated by arrows: an arrow behind the gesture indicates a decrease of temporal extent of the gesture; An arrow before the gesture indicates a shift of the gesture to the left, i.e. an increase in overlap with other gestures. The horizontal extension of the boxes represents the time interval of gestural activation. Association lines are indicated by vertical dotted (case syllable boundary) or dashed (all other cases) lines. Time intervals of consonantal obstructions (closures) are indicated by shaded areas within the boxes of gestural activation. Time intervals of gottal or velic closing movements following an opening gesture are indicated by dashed lined boxes if necessary.

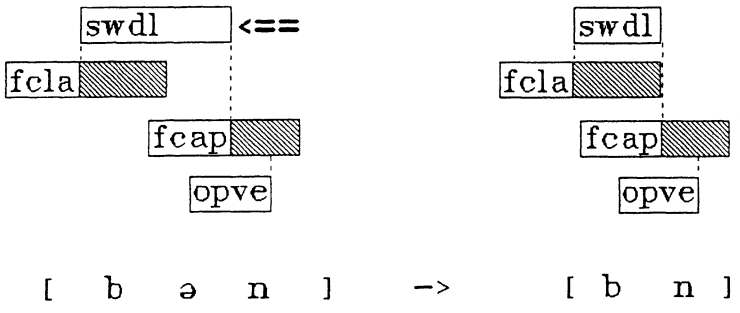


Figure 5a Gestural alteration process for elision of [ə] in "eben".

Elision of [ə] in "eben" ([bən]->[bn]) is reached by a decrease in temporal extent of the dorsal labial schwa-forming gesture {swdl} (fig. 5a). According to the gestural phasing rules this process also leads to an increase in temporal overlap of the labial and apical full-closing gestures {fcla} and {fcap}. The segmental elision occurs if the closure intervals of both closing gestures (shaded areas in both gestural activation intervals) overlap. It should be noted that segmental elision of [ə] occurs without a full reduction of the temporal extent of the schwa-gesture. The schwa-gesture still exists in the reduced form. This gesture is only hidden by the occlusions of the consonantal gestures.

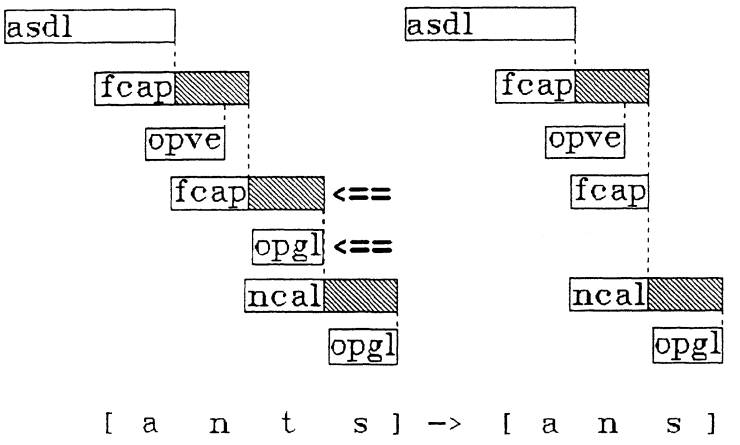


Figure 5b Gestural alteration process for elision of [t] in "Glanz".

Elision of [t] in "Glanz" ([nts]->[ns]) is reached by decrease in temporal extent of the apical full-closing gesture {fcap} and glottal opening gesture {opgl} (fig. 5b). The glottal opening gesture may remain unreduced if the reduction of temporal extent of the apical full-closing gesture does not lead to a shift of the temporal location of the glottal opening gesture.

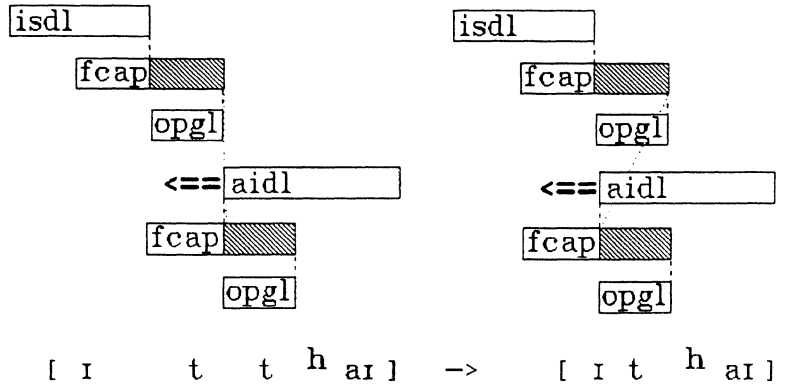


Figure 5c Gestural alteration process for reduction of double consonants in "mitteilen".

Reduction of double consonants in "mitteilen" ([tt]->[t]) is reached by increase in temporal overlap of the first and the second syllable, i.e. by increase in overlap of the vocalic gestures of the first and second syllable: the dorsal-labial short /i/-forming gesture {isdl} and the dorsal-labial /ai/-forming gesture {aidl} (fig. 5c). According to the gestural phasing rules this leads to a complete temporal overlap of the apical full-closing gestures {fcap} and their appertaining glottal opening gestures {opgl}, i.e. a complete temporal overlap of the initial consonantal closing gestures and their appertaining opening gestures of the second syllable with the final consonantal gestures of the first syllable.

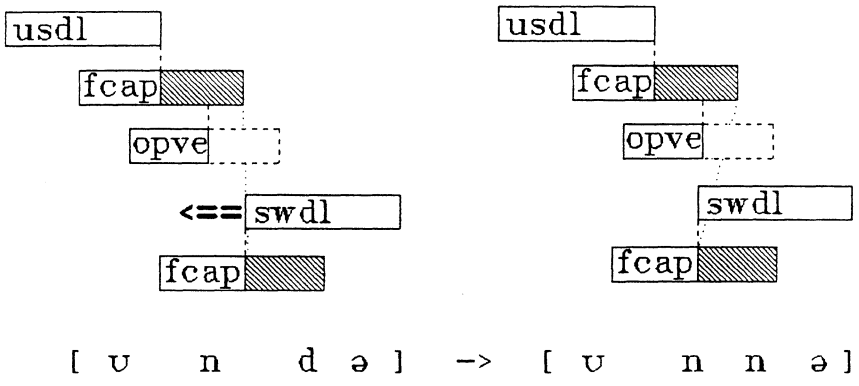


Figure 5d Gestural alteration process for progressive assimilation of nasality in "Bundes".

Progressive assimilation of nasality in "Bundes" ([nd]->[nn]) is reached by increase in temporal overlap of the first and the second syllable, i.e. by increase in overlap of the vocalic gestures of the first and second syllable: the dorsal-labial short /u/-forming gesture {usdl} and the dorsal-labial schwa-forming gesture {swdl} (fig. 5d). According to the gestural phasing rules this leads to a complete temporal overlap of the apical full-closing gestures {fcap}, i.e. a complete temporal overlap of the initial consonantal closing gesture of the second syllable with the final consonantal gestures of the first syllable.

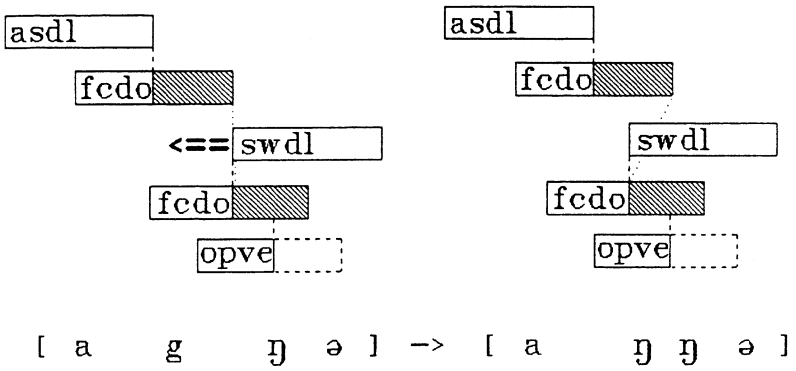


Figure 5e Gestural alteration process for regressive assimilation of nasality in "Agnes".

Regressive assimilation of nasality in "Agnes" ([gŋ]->[ŋŋ]) is reached by increase in temporal overlap of the first and the second syllable, i.e. by increase in overlap of the vocalic gestures of the first and second syllable: the dorsal-labial short /a/-forming gesture {asdl} and the dorsal-labial schwa-forming gesture {swdl} (fig. 5e). According to the gestural phasing rules this leads to a complete temporal overlap of the dorsal full-closing gestures {fcdo}, i.e. a complete temporal overlap of the initial consonantal closing gestures of the second syllable with the final consonantal gestures of the first syllable.

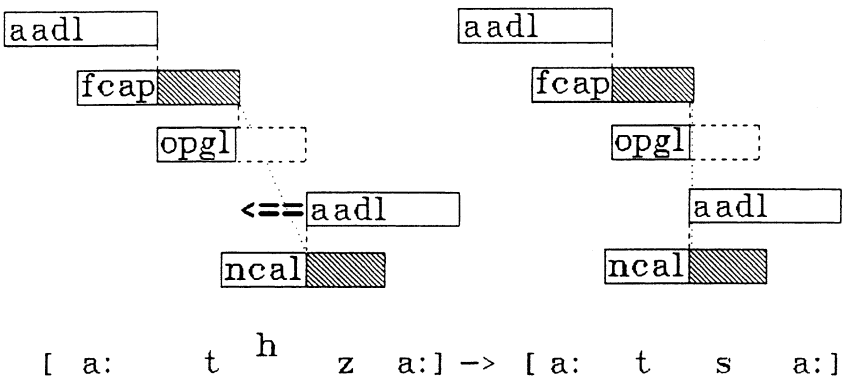


Figure 5f Gestural alteration process for progressive assimilation of voicelessness in "ratsam".

Progressive assimilation of voicelessness in "ratsam" ([tz]->[ts]) is reached by increase in temporal overlap of the first and the second syllable, i.e. by increase in overlap of the vocalic gestures of the first and second syllable: both dorsal-labial long /a/-forming gestures {aadl} (fig. 5f). But in this case the temporal overlap is not robust as in the above cases of overlap of syllables. According to the gestural phasing rules this temporal overlap leads to a partial temporal overlap of the alveolar near-closing gesture {ncal} with the apical full-closing gesture {fcap}, i.e. a partial temporal overlap of the initial consonantal closing gestures of the second syllable with the final consonantal gestures of the first syllable.

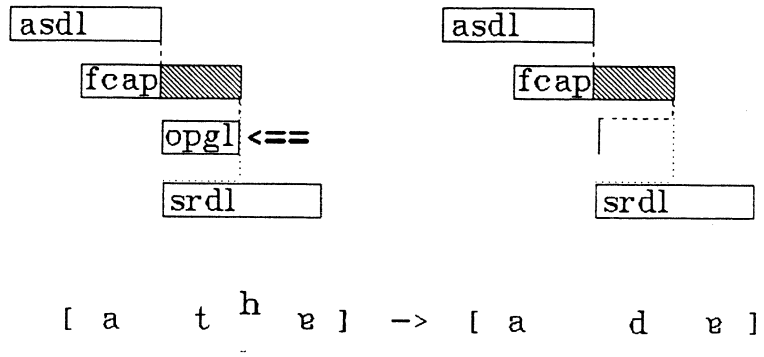


Figure 5g Gestural alteration process for sonorization in "hat er".

Sonorization (i.e. assimilation of voice) in "hat er" ([t^h]->[d]) is reached by decrease in temporal extent of the glottal opening gesture {opgl} (fig. 5g). This decrease need not lead to a total reduction of the glottal opening gesture. The segmental change is also reached if a rudimentary glottal opening gesture remains. If the phonological gestural concept is extended by introducing glottal closing gestures in order to ensure the occurrence of glottal vibration (i.e. of voicing) this process is equivalent to an increase in overlap of the gottal opening gesture and the glottal closing gesture of the vocalic portion of the second syllable. In the present gestural concept voicing results from glottal underspecification. But an extension of our gestural approach by introducing glottal and velic closing gestures in the case of sonorants and obstruents has been suggested (Geuman and Kröger 1995).

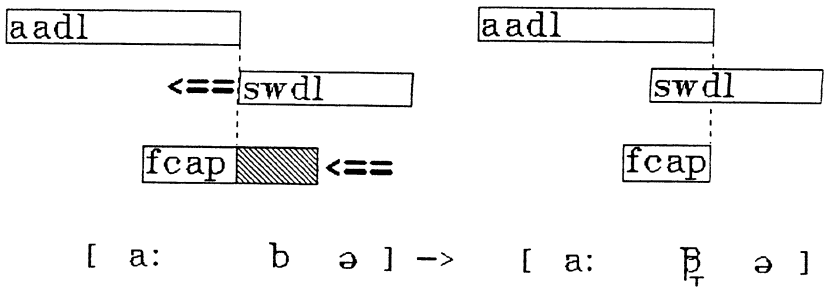


Figure 5h Gestural alteration process for reduction of degree of opening in "ich habe".

Reduction of degree of opening in "ich habe" ([b] -> [β̞]) is reached by a decrease in temporal extent of the apical full-closing gesture {fcap} (fig. 5h). This decrease quantitatively leads to a release phase value of around 180 degrees. Thus the decrease of temporal extent leads to an omission of the consonantal closure interval. Since this makes the transition portions of the vocalic gestures of the second syllable audible, a rearrangement of the timing of the vocalic gestures is necessary. Thus the increase in overlap of the dorsal-labial schwa-forming gesture {swdl} occurrs together with the dorsal labial long /a/-forming gesture {aadl}.
For all cases given above, discrete segmental changes can be generated by continuous gestural alteration processes, which confirms a main hypothesis of Articulatory Phonology. But three categories of segmental changes remain which cannot be generated by a continuous gestural alteration process even if the identification of the hypothetical gestural process is no problem: the regressive assimilation of manner and the progressive and regressive assimilation of place. We will focus here on regressive assimilation of place. In the case of the examples "mit mein(e)m", "mit jed(e)m", and "mit fett(e)m" the gestural alteration process is easy to identify: decrease of temporal extent of the schwa-gesture of the last syllable. After the elision of [ə] we expect the segmental change associated with regressive assimilation of place: the change [nm]->[mm], [dm]->[bm], and [tm]->[pm], i.e. a change from "apical" to "labial" for the last but one consonant. But the transcriptions of the generated reduced forms do not indicate this change in all cases. In many transcriptions the changes [nm]->[nn], [dm]->[dn], and [tm]->[tn] occur. In order to clarify these findings we performed quantitative listening tests.

3.2 Listening tests for "mit meinem", "mit jedem", and "mit fettem"

Since so far the transcriptions were done by a single person, we performed a quatitative listening test. This seems to be important especially in the case of regressive assimilation of place for the forms "mit meinem", "mit jedem", and "mit fettem". As stated above, the trained phonetician does not perceive clearly regressive assimilation of place in all cases if a decrease of temporal extent of the schwa-gesture of the last syllable is introduced (fig. 6a). It is our hypothesis, that a further gestural process must be introduced to Articulatory Phonology: *swap of the gesture-executing articulator*. In the case of our examples this is a swap of the apical full-closing gesture of the last syllable for a labial full-closing gesture (fig. 6b). This is a discrete gestural process and will be labelled *gestural or articulatory reorganisation.*

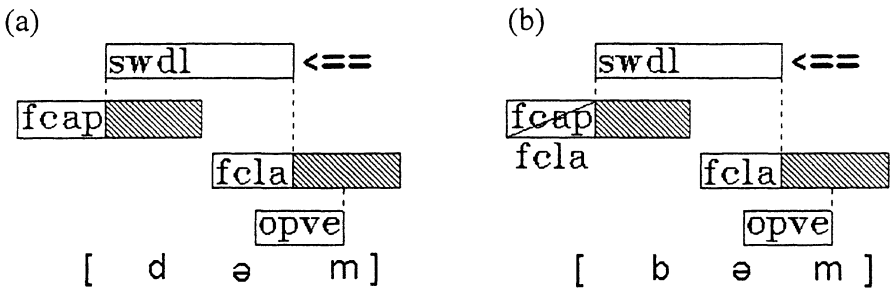
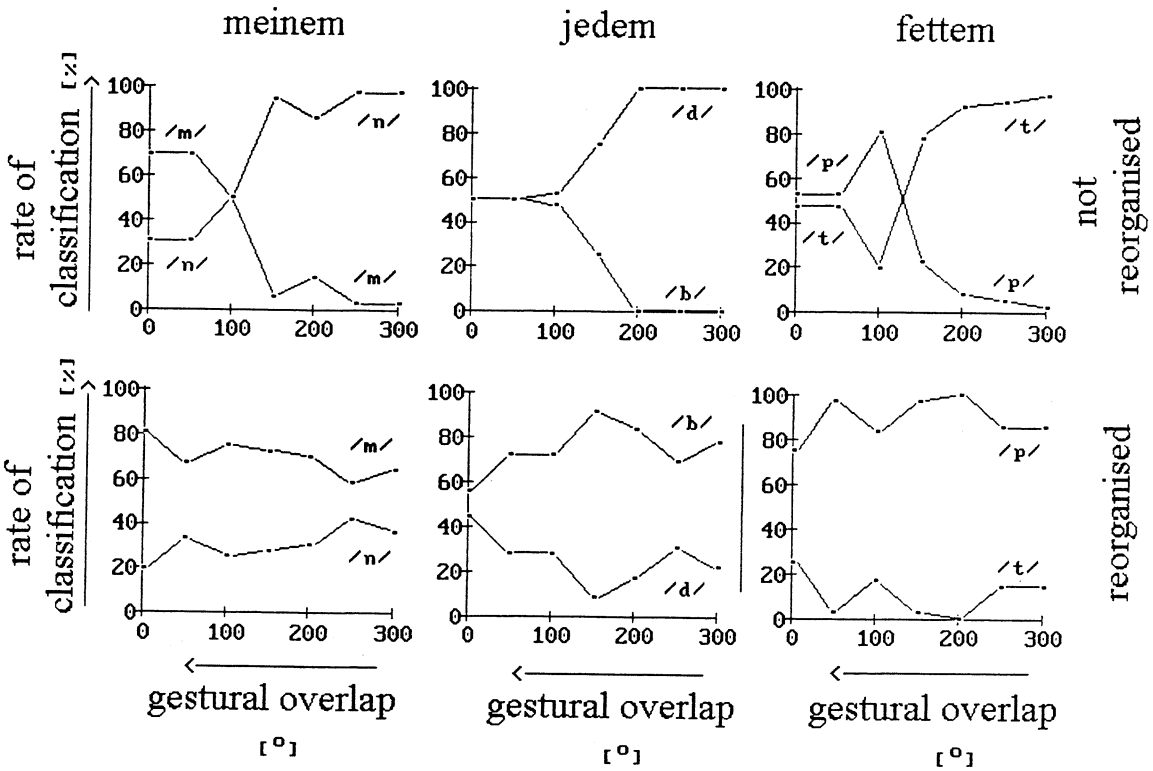


Figure 6 Continuous gestural alteration process (arrow) **(a)** without and **(b)** after reorganization for "mit jed(e)m".

In order to show that a continuous gestural alteration process is not sufficient to generate regressive assimilation of place, we performed quantitative listening tests. We started with six forms, i.e. the reorganised and not reorganised forms of "mit meinem", "mit jedem", and "mit fettem". For each form we performed the continuous gestural alteration process indicated in figure 6, i.e. decrease of temporal extent of the schwa-gesture of the last syllable. This process leads to an increasing degree of gestural overlap of the initial and final constriction-forming gestures of this syllable. Seven stimuli covering the whole range of the continuous gestural alteration process from no overlap to maximum gestural overlap of the constriction-forming gestures were generated for each of the six forms. The total of 42 stimuli was presented in random order with three repetitions of each stimulus to a group of 12 listeners (students of phonetics). Two tests were performed using the same stimuli but with different tasks: classification of the last but one consonant (test 1) and classification of the last consonant (test 2) to the category "labial" or "apical".

The rates of classification are given in figure 7. In the case of no reorganisation we find a tendency from apical to labial with increasing gestural overlap. But this tendency for regressive assimilation of place is not robust. Even in the case of full gestural overlap (left side of the diagrams) we find a maximum classification rate for "labial" of only around 50% to 70% for these three forms. But in the case of test 2 we find a robust tendency from "labial" to "apical", i.e. a tendency *against* regressive assimilation of place. Consequently, regressive assimilation of place has not been perceived clearly in the case of the non-reorganised forms. On the other hand, in the case of reorganisation a high rate of "labial" is perceived, regardless of the degree of gestural overlap. This result was to be expected since in the case of reorganisation we have only labial closing gestures on the side of articulation. But it excludes any immanent tendency towards "apical". So gestural reorganisation is necessary in the case of these forms in order to generate regressive assimilation of place.

(a)



(b)

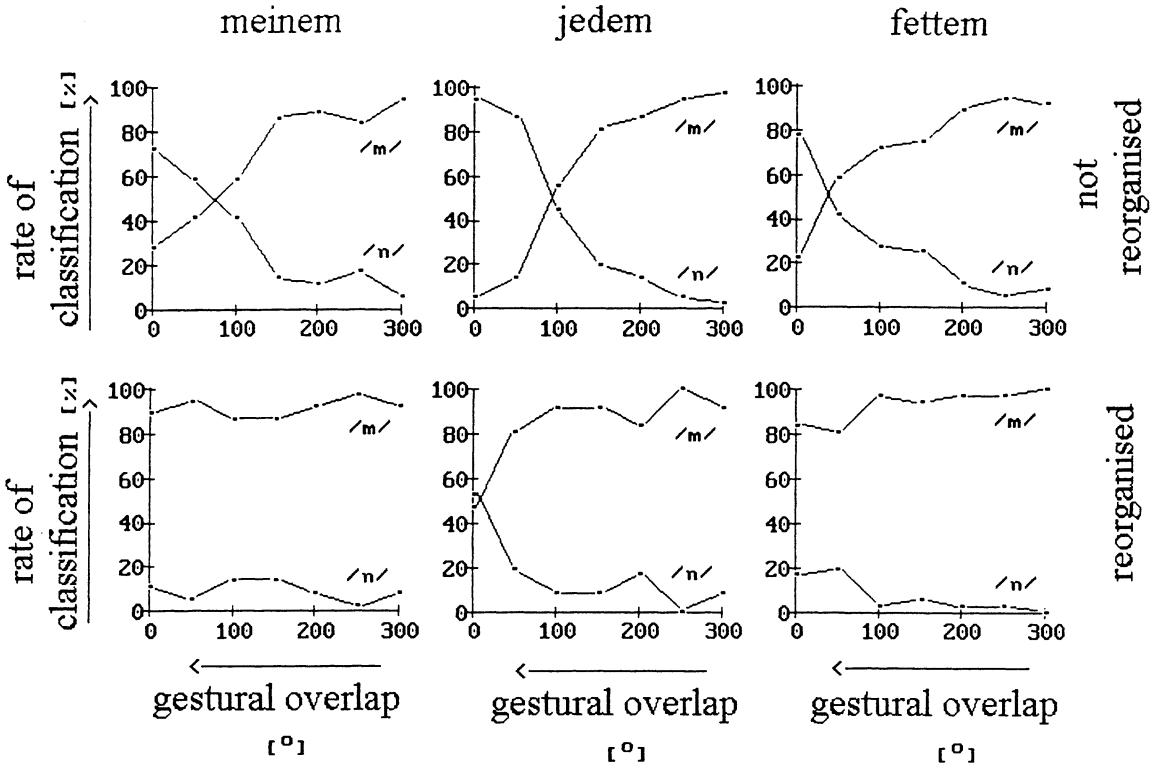


Figure 7 Rates of classification as function of gestural overlap (association phase values; increasing overlap from the right to the left side) for classification of the last but one (**a**; test 1) and the last (**b**; test 2) consonant for the three forms in the reorganised and the not reorganised case.

4 A gestural theory of reduction

It has been elucidated so far that different gestural processes are involved in reduction: two continuous gestural processes - i.e. increase in gestural overlap of gestures and decrease in temporal extent of a gesture - and one discrete gestural process: gestural reorganization. We assume one unique underlying driving force for all continuous and discrete gestural processes in reduction: *minimization of articulatory effort* (Lindblom 1990).

In our concrete gestural approach articulatory effort can be expressed quantitatively. Articulatory effort of an utterance can be defined as the sum-total of articulatory effort of all gestures occurring within an utterance. And effort of each single gesture is proportional to the duration of the appertaining gestural activation interval. We will find that this simple quantitative model of articulatory effort is capable of motivating many of the gestural processes in reduction.

In a more complex quantitative model of articulatory effort we can differentiate between effort of movement, i.e. effort of the transition portion, and effort of the gestural hold portion. The effort of the transition portion increases with increase in length of its time interval and with the distance of the target position to the initial articulator position - i.e. increases with the gestural movement amplitude - and increases with decrease in eigenperiod - i.e. increases with increase in strength of gestural activation (Kröger et al. 1995 and Kröger 1997).

In order to show that all gestural processes involved in reduction lead to a decrease in articulatory effort, we have to analyze the change in articulatory effort of an utterance (a word) for each gestural alteration process.
(1) Decrease of the temporal extent of a gesture: In this case articulatory effort decreases assuming that all other gestural parameters are constant. This follows directly from our quantitative expression for articulatory effort.
(2) Increase in temporal overlap of two gestures: (2a) If two gestures act on the same articulator, increase in overlap leads to a reduction in temporal extent of the preceding gesture. In this case gestural overlap is modelled by separating the time interval of overlap continuously for both gestures. In the first part of the interval of overlap mainly the first (or preceding) gesture remains active while in the second interval of overlap mainly the second (or following) gesture is active. This leads to a shortening of the temporal extent of both gestures. In our simple quantitative gestural approach (Kröger 1993) we introduced a dominance of the following gesture if two gestures overlap on the same articulator. This leads to a shortening of the temporal extent of the first (or preceding) gesture. But in both cases, we get a decrease in articulatory effort. This holds for all cases, i.e. for increasing the temporal overlap of tract-shaping gestures (i.e. increase in overlap of syllable cores; see chapter 3.1), for increasing the temporal overlap of constriction-forming gestures acting on the same articulator and of opening gestures acting on the same articulator. (2b) If two gestures act on different articulators the decrease in articulatory effort from other sources. Here we must differentiate three main groups (chapter 3.1): (A) Increase in overlap of constriction-forming gestures occurring in a syllable initial consonant cluster with those occurring in a syllable final consonant cluster. This process presumes the decrease of temporal extent of the tract-shaping gesture of this syllable (e.g. in the case of elision of [ə], example "eben" chapter 3.1). (B) Increase in overlap of constriction-forming gestures occurring in a syllable final consonant cluster with those of the syllable initial consonant cluster of the following syllable. This case occurs only if the constriction-forming gestures act on the same articulator (e.g. for reduction of double consonants in "mitteilen", for regressive assimilation of nasality in "Agnes" or for sonorization in "ratsam", chapter 3.1). For constriction-forming gestures acting on different articulators this case leads to a decrease in articulatory effort, if overlap of tract-shaping gestures is reached leading to a decrease in temporal extent of these gestures. This condition is satisfied if tract-shaping gestures overlap more strongly. (C) Increase in overlap of constriction-forming gestures occurring in the syllable initial or final consonant cluster. This case occurs only if the constriction-forming gestures act on the same articulator (e.g. for elision of [t] in "Glanz", chapter 3.1). For constriction-forming gestures acting on different articulators this case leads to a decrease in articulatory effort if overlap of tract-shaping gestures is reached which leads to a decrease in temporal extent of these gestures.
(3) Swap of gesture-executing articulator: This discrete gestural alteration process reduces the articulatory effort of an utterance in definite cases: (3a) Swap of constriction-forming gestures from tongue tip to lips or tongue body. In these cases we have a swap from the apical to the labial or dorsal tier. Since the tract-shaping gestures (i.e. the vocalic gestures) are always active on these tiers, this swap leads to an increase in overlap of gestures on the labial or dorsal tiers and subsequently to a reduction in articulatory effort. It is important to mention that the condition of swap from apical to labial or dorsal covers all cases of assimilation of place occurring in German (see chapter 3.2). So only shifts from apical to labial or from apical to dorsal occur but not vice versa. (3b) Furthermore it should be mentioned that the occurrence of glottalization and glottal stops (Kohler 1994) can be interpreted as a swap of a closing gesture

from the velum to the glottis in the gestural framework (e.g. in "Stund(e)n" [ndn] -> [n?n] or as glottalization: [ndn] -> [nn̰n]). The glottal tier exhibits phonatory gestures without temporal gap: Closing gestures with medium adductive force produce voicing and opening gestures produce voicelessness (In the case of our simple model voicing results from gestural underspecification on the glottal tier). Consequently this swap to the glottal tier leads to an increase in gestural overlap of the (former velic) closing gesture with phonatory gestures. Furthermore, in the more complex model of articulatory effort we can assume that a swap of a closing gesture from velum to glottis leads to a decrease in gestural movement amplitude and thus to a decrease in effort of movement.

5 Conclusions

We have illustrated some advantages of Articulatory Phonology. A close relationship between phonological description and its phonetic realization is reached. The gesture as central unit of this theory can be seen as a phonetic unit (unit of action, goal-directed articulator movement) as well as a phonological unit (unit of phonological contrast). Articulatory Phonology is able to ascribe a lot of different discrete segmental changes (elisions and assimilations) to simple continuous gestural alteration processes: decrease in temporal extent of a gesture and increase in temporal overlap of gestures. But some segmental changes (e.g. assimilation of place in German) can be realized only by the discrete gestural process called "swap of gesture-executing articulator" or "gestural reorganization". It has been emphasized that the underlying driving force for all gestural processes mentioned here is the minimization of articulatory effort. In future work we will try to establish a complete rule system for the specification of concrete gestural processes in given utterances. This, however, will make it necessary to incorporate both articulation and perception into Articulatory Phonology.

6 Literature

Browman, C.P., Goldstein, L. (1989): "Articulatory gestures as phonological units", *Phonology* **6**, 201-251.

Browman, C.P., Goldstein, L. (1990): "Tiers in articulatory phonology, with some implications for casual speech", in: J. Kingston, M.E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (Cambridge University Press, Cambridge), S. 341-376. Also in: *Haskins Laboratories Status Report on Speech Research* **SR-92** (1987), 1-30.

Browman, C.P. (1991): "Consonants and vowels: Overlapping gestural organization", *Proceedings of the XIIth International Congress of Phonetic Sciences*, Vol.1, 379-383.

Browman, C.P., Goldstein, L. (1992): "Articulatory phonology: An overview", *Phonetica* **49**, 155-180.

Fujimura, O. (1981): "Temporal organization of articulatory movements as a multidimensional phrasal structure", *Phonetica* **38**, 66-83.

Fujimura, O. (1986): "Relative invariance of articulatory movements: An iceberg model", in: J.S. Perkell, D.H. Klatt (Eds.), *Invariance and Variability in Speech Processes* (Lawrence Erlbaum, Hillsdale, New Jersey), 226-242.

Fujimura, O. (1992): "Phonology and phonetics - A syllable-based model of articulatory organization", *Journal of the Acoustical Society of Japan (E)*, **13**, 39-48

Geumann, A., Kröger, B.J. (1995): "Some implications for gestural underspecification as a result of the analysis of German /t/-assimilation", *Proceedings of the 13th International Congress of Phonetic Sciences,* Vol. 3, 374-377.

Kohler, K.J. (1994): "Glottal stops and glottalization in German", *Phonetica* **51**, 39-51.

Kohler, K.J. (1995): *Einführung in die Phonetik des Deutschen* (Erich Schmidt Verlag, Berlin), 2. edition.

Kröger, B.J. (1993): "A gestural production model and its application to reduction in German", *Phonetica* **50**, 213-233.

Kröger, B.J., Schröder, G. Opgen-Rhein, C. (1995): "A gesture-based dynamic model describing articulatory movement data", *Journal of the Acoustical Society of America* **98,** 1878-1889.

Kröger, B.J. (1997): "Ein quantitatives Konzept des artikulatorischen Aufwandes", in: K. Fellbaum (Ed.), *Elektronische Sprachsignalverarbeitung. Studientexte zur Sprachkommunikation* **14**, 248-255.

Liberman, A.M., Mattingly, I.G. (1985): "The motor theory of speech perception revised", *Cognition* **21,** 1-36.

Lindblom, B. (1990): "Explaining phonetic variation: A sketch of the H and H theory", in: W.J. Hardcastle, A. Marchal (Eds.), *Speech Production and Speech Modelling* (Kluwer Academic Press, Dordrecht), 403-440.

Löfqvist, A., Yoshioka, H. (1984): "Intrasegmental timing: Laryngeal-oral coordination in voiceless consonant production", *Speech Communication* **3**, 279-289.

Öhman, S.E.G. (1967): "Numerical model of coarticulation", *Journal of the Acoustical Society of America* **41**, 310-320.

Saltzman, E.L., Munhall, K.G. (1989): "A dynamical approach to gestural patterning in speech production", *Ecological psychology* **1**, 333-382.

Variability in articulation and timing in connected speech of different style

Lioba Faust
Institut für Kommunikationsforschung und Phonetik (IKP)
Poppelsdorfer Allee 47, 53115 Bonn, Germany
Phone: +49-228-735641 E-Mail: lfa@ikp.uni-bonn.de

Abstract

In this study, variation in speech utterances of the same speakers in different communication situations was examined. A corpus of seven dialogues was recorded each in four different types, produced by fourteen speakers. From the corpus, utterances by two speakers were selected for a listening experiment in order to find out to what extent listeners are able to identify spontaneous and read utterances. Variability in speech was analysed with regard to the speakers and to the different dialogue types. The global structure of speech production was observed for timing processes: speech rate, articulation rate and pauses were measured. Speech production was analysed as deletions, substitutions and insertions of sounds. The results revealed that variation is strongly speaker-dependent. However, casual speech tends to be less carefully articulated than read speech. The two speakers selected for the listening experiment showed contrasting articulatory patterns with corresponding classifications by the listeners. The results illustrate that speaking styles have to be regarded as productions of different communication situations with each one showing its own speaker-dependent articulatory pattern.

1. Introduction

It is well known that speech varies to a large extent depending on different speakers. It is also a fact that different communication situations have an influence on speech utterances. Research in phonetics and speech engineering has been focusing on the variability in connected speech collected not only under laboratory conditions but also in real-life situations. Up to now, speech engineering - and especially speech recognition - for the most part analysed read speech recorded under lab conditions. As in everyday use read speech can differ largely from speech uttered spontaneously, it is no longer sufficient to be restricted to read speech. The more research is performed, the less satisfying the rough distinction between read and spontaneous speech becomes. Considering speech utterances from different communication situations, "speaking style" was defined as a promising term. In order to collect real-life speech utterances from different speaking styles, a couple of techniques were developed, e.g. the Wizard-of-Oz-experiment, interview talks etc. Generally, the term "speaking styles" was used to characterize casual or informal speech, careful, formal or clear speech and read speech (Eskénazi 1993, Blaauw 1994, Beckman 1995).

Of course, the collection of speech data for acoustic-phonetic analysis is rather difficult and, to some extent, has to be performed under lab conditions. However, it is clear that the distinction between spontaneous and read cannot be sufficient to characterize speech utterances. Before determining utterances as belonging to a certain speaking style or type of speech, it is interesting to analyse their phonetic variability.

As a starting point, it is important to look at the speech we are used to in real life. Indeed, we find a lot of examples for speech uttered spontaneously, and less examples for speech read out aloud. In table 1 some examples are presented.

Table 1: Examples of communication situations

Communication situations and "speaking styles" of everyday conversation	
➤ chatting ➤ talking with friends ➤ speaking to the boss ➤ explaining something to foreigners ➤ telling jokes ➤ interview talks ➤ talking to a child ➤ talking to hearing-impaired persons ➤ discussing	➤ professional news-reading ➤ reading fairy tales to a child ➤ reading a letter/an article in a newspaper ➤ reading to a blind person ➤ professional reading of literature
phrased by the speaker	composed by another person
"spontaneous speech"	"read speech"

From the examples we learn that different "speaking styles" are found almost as often as there are different communication situations. The distinction between spontaneous and read speech is only related to the fact that spontaneous speech is phrased by the speaker himself or herself in the moment of speaking and that read speech is written down and, at least ordinarily, composed by another speaker. We cannot yet conclude that there are great phonetic differences between speech uttered spontaneously or read out, but we may perceptually observe variation in different communication situations and for the same speakers.

The present study attempts to give an answer to the following questions:

1. What about the *range of variability* within and between different speech styles?
2. Are there *crucial differences* between read speech and speech uttered spontaneously?
3. What happens to the *production of speech* when a speaker changes to another speaking style?

2. Description of the experiment

2.1 Design and performance of the speech corpus

In order to get rather natural utterances in real-life conversation, dialogues were recorded in four different types: fourteen speakers (eleven female, three male) between 21 and 27 years of age participated in seven dialogue recordings of German with each one being performed in the following different speech types:

1. Recording of a *totally free conversation* without any instructions. The speakers did not know that they were recorded and the communication situation was relaxed. The resulting speech was casual.

2. Recording of a *time-scheduling negotiation dialogue* using a formal mode of address. The resulting speech was supposed to be less casual than in the first dialogue.

3. The task for the speakers was the same as in 2., but the recording conditions were controlled, i.e. the speakers had to press a button before they started speaking, and they could not speak simultaneously.

4. The last recording was a re-reading of the transcribed utterances of the second dialogue type. Hesitations, word repetitions and repairs that had been produced and transcribed were generally dropped for the copy to be re-read. The most important issue for the re-reading copy was to preserve the dialogue structure and, in general, the grammatical structures of the utterances.

The speakers were sitting in the institute's speech laboratory. The communication was performed using headset microphones (Sennheiser HMD 414-6) for the dialogue types 2 and 3 and room microphones (Neumann KM 140) for the casual and the read conversation. Speech was digitally recorded on two separate channels.

2.2 The method of experimentation

2.2.1 Listening experiment

Ten phonetically educated and twelve naïve listeners (beginning students) were asked to classify the selected utterances as "spontaneous" or "read" in a forced choice task. Furthermore, the listeners had to rate the degree of reliability for their decision on a five-point scale from "very safe" to "very unsafe". Moreover, the essential linguistic or phonetic features underlying the listeners' decisions had to be specified. The given features were: syntactic structure, speech fluency, repairs, articulation, intonation, and speech rate.

For the listening experiment the selection of utterances out of the corpus was done on the following conditions: One female speaker was selected whose spontaneous and read speech patterns appeared to differ very much. A second female speaker from another dialogue was selected whose speech patterns seemed to be quite similar in both dialogue types. From each speaker, three utterances were selected from the dialogue types 1, 2 and 4. The utterances were of phrasal length, grammatically sentence-like, with different intonational patterns, and they contained a certain amount of pauses, hesitations and repairs. From the careful spontaneous and the read versions, identical utterances were collected. The utterances were presented in random order, and each utterance was played twice.

2.2.2 Proceeding for segmental analyses

All dialogue utterances were transcribed literally, segmented and labelled manually and marked with phrasal accents. The labelling was performed by the author and two other trained phoneticians.

It was decided not to be restricted to segmental analyses on the sound level, but to examine as well the global structure of speech production, i.e. **timing structure**. Timing was measured as

1. *speech rate* including any kind of pauses, hesitations or lengthenings,
2. *articulation rate*, i.e. fluently produced utterances, and
3. *"interruptions"*, i.e. pauses and sound-lengthenings, which were also measured according to their phrasal position.

Timing structure was decided to be measured as number of sounds/s (and, concerning speech rate, labels/s). The usual manner in measuring timing structure as the number of syllables/s (Kohler et al. 1981) was rejected for two reasons: 1. Especially in the first dialogue type, the number of syllables was very difficult to determine as non-accented syllables were, at least for some speakers, to a large extent contracted and, thus, deleted. Therefore, it was more obvious to examine sounds instead of syllables. 2. The analysis of sounds had to be performed anyway for the measuring of articulation.

Speech rate was measured as the number of labels/s, articulation rate as the number of fluently produced sounds/s where "interruptions" had been excluded. "Interruptions" were classified into breathing, silent and filled pauses, and sound-lengthenings. Each kind of interruption was labelled by a certain label symbol in order to make it easy to exclude them from the measuring of articulation rate and to examine them separately. Sound-lengthenings were also observed in different positions: 1. utterance-final, 2. phrase-final followed by a pause, 3. phrase-final, 4. interrupting (i.e. within a phrase) followed by a pause and 5. interrupting. The amount of pauses was examined as occurring non-interrupting (i.e. corresponding to a syntactic boundary) or interrupting.

Sounds were marked as lengthened using the following method: The average sound length of each sound category was measured for each speaker and each dialogue type, then calculated for all speakers. From the sounds that had been manually labelled as lengthened the lowest value of a sound category was noted and divided by the calculated average for each sound category. Thus, a coefficient resulted of which the average for all sound categories was calculated, and, afterwards, for all dialogue types. The resulting coefficient was 2.5. Multiplying the sound length of all labelled sounds by 2.5, the resulting values were compared to the real produced sound length by a computer program. Longer sounds were automatically labelled as lengthened, shorter sounds were not.

Measuring **articulation**, reductions were examined as the number of sound deletions, substitutions and insertions compared to the citation form of the corresponding word which was taken as reference. Furthermore, segmental reductions were measured according to sound categories (plosives, fricatives, nasals, vowels, glides, liquids).

The first step for the examination of speech production was to carry out the reference phonetic transcription using a computer program that converted the literal transcription into the citation form (P-TRA, created at IKP, Stock 1992).

The citation form of the words then was taken as the reference for the measurement of reductions. The comparison of the symbol strings was performed automatically by applying a window that included a variable number of symbols to be compared. Finally, the total amount of sounds produced by a speaker in a dialogue type corresponded to the labelling string. This labelling string was compared to the reference string. Thus, missing sounds in relation to the citation form were defined as deletions, sounds that were different from the citation form were defined as substitutions. Sound substitutions according to sound categories were analysed manually. Sounds left in the labelling string compared to the citation form were examined as insertions.

3. Results

A summary of results will be presented concerning the correlation between the different speech styles and dialogue types, respectively. Concerning the listening experiment, the results for the two selected speakers will be presented separately.

3.1 Speaker-related results

3.1.1 Listening experiment

For the listening experiment, identical utterances from dialogue type 2 (careful spontaneous speech) and 4 (read) were selected, and, in order to examine to what extent casual utterances could be classified as spontaneous, utterances from dialogue type 1 were presented to the listeners as well.

Table 2: Correct classification of utterances (percentage for 22 listeners)

Speaker	dialogue type 1 ("casual")	dialogue type 2 ("careful")	dialogue type 4 (read)
A	97.0%	93.9%	62.1%
B	90.6%	62.1%	43.9%

Table 2 illustrates the correct classification of utterances, i.e. utterances of dialogue type 1 and 2 were identified as spontaneous, and those of type 4 as read. For both speakers, the casual utterances were correctly identified in almost all cases. This by far clear result for the casual style may be due to the restricted acoustic conditions and to the informal content of the dialogue utterances.

The careful style requires a closer look at the individual speakers: The utterances of speaker A are to a large extent still classified as spontaneous, whereas for speaker B there occur misclassifications, i.e. more than a third of the listeners classifies the utterances as read. For the read speech style it is interesting to note that the utterances of speaker A are mostly classified as read, but that for speaker B the majority of listeners classifies the read utterances as spontaneous. From these results it may be concluded that casual speech is identified as spontaneous, whereas the identification of careful and read speech seems to be dependent on the speaker.

As to the reliability of listeners' decisions, listeners are generally "very safe" or "rather safe" about reaching a correct decision.

Table 3: Correct classification due to *linguistic features* (Most frequent answers by trained listeners)

Speaker	dialogue type 1 ("casual")		dialogue type 2 ("careful")		dialogue type 4 (read)	
A	articulation speech rate	63% 63%	fluency	73%	intonation	67%
B	intonation fluency	48% 44%	syntax	59%	intonation	75%

For the correct decisions, the distribution of the phonetic or linguistic features was examined. The most frequent answers by trained listeners are illustrated in table 3. Concerning the casual style, all features are mentioned more or less frequently by the listeners. In the careful style, "fluency" is most often mentioned for speaker A, for speaker B it is "syntactic structure". For the read speech style "intonation" is the convincing feature for the listeners to classify the utterances of both speakers correctly.

The results from the listening experiment show that especially for casual speech there are different linguistic features that dominate the perceptual impression of the listeners leading them to a correct classification. From this we learn that it is necessary to have a look at the "patterns" in speech in order to understand the listeners' decisions.

3.1.2 Timing structure

Since the analysis of speech rate contains all "interrupting parts" of the utterances, speech rate does not reveal satisfying results. To get an impression of the timing structure in the utterances of the two selected speakers, it is most interesting to examine articulation rate for the three dialogue types that had been presented to the listeners. The values of speech rate and interruptions for the whole corpus will be shown in section 3.2.1.

Table 4: Articulation rate (number of sounds/s)

Speaker	casual (1)	careful (2)	read (4)	development
A	16.9	14.6	13.7	↘
B	13.0	16.3	16.3	↗

Table 4 illustrates articulation rate as number of sounds/s. Comparing the results for the two speakers, the contrast is immediately perceivable: Whereas for speaker A the highest articulation rate occurs in the casual speech type, for speaker B it occurs in the read dialogue type. For both speakers the development from the casual towards the read speech type is obviously contrasting, i.e. for speaker A it is decreasing, for speaker B it is increasing. From these results we can conclude that speaker A speaks slower when reading than when speaking spontaneously, and, in contrast to her, speaker B articulates faster in read than in casual speech. Considering the total corpus of speakers, another important item has to be mentioned: The articulation rate of speaker A in casual speech is the highest compared to all speakers, speaker B reaches in casual speech the lowest, in read speech the highest rate in the total corpus.

These contrasting results for articulation rate may already explain the decisions of the listeners in the listening experiment.

3.1.3 Speech production

The results for speech production concerning the two selected speakers will also be restricted to the three dialogue types that were presented in the listening experiment.

Table 5: Deletions (missing sounds related to the total amount of sounds produced; speaker-specific)

Speaker	dialogue type 1 ("casual")	dialogue type 2 ("careful")	dialogue type 4 (read)	development
A	17.6%	8.5%	4.5%	↘
B	16.4%	8.0%	10.4%	↘↗

Table 6: Substitutions (related to the total amount of sounds produced; speaker-specific)

Speaker	dialogue type 1 ("casual")	dialogue type 2 ("careful")	dialogue type 4 (read)	development
A	13.2%	6.7%	4.9%	↘
B	10.2%	6.8%	9.9%	↘↗

For both speakers, the examination of reductions yields very clear results. Whereas the amount of reductions decreases for speaker A from casual speech towards read speech, the development for speaker B is vice versa. Table 5 illustrates the amount of deleted sounds compared to the total amount of sounds produced in this speech type.

In casual speech it is interesting to note that the amount of deletions for both speakers is similar, i.e. they do not articulate very carefully. In dialogue type 2 both speakers still reach a similar amount of reductions, and much less than in casual speech. When reading, speaker B increases the amount of deletions, whereas speaker A decreases even further.

For the articulation rate, the results are corresponding: They illustrate that speaker A articulates more slowly and more carefully the less "spontaneous" the situation, whereas for speaker B reading is not very different from speaking spontaneously as she articulates faster and just as little careful as in casual speech.

We can conclude that the manner of articulation is different for the two speakers according to the communication situation.

The results for substituted sounds (Table 6) show a similar pattern as for deletions and, thus, underline the development of speech production for the two speakers.

The conclusion that can be drawn from the experiments is as follows: For each speaker, the results for speech production, articulation rate and perceptual classification correspond. The less spontaneous the communication situation, the more careful the articulatory pattern of speaker A and the slower the articulation. Her utterances generally are correctly classified. The articulatory pattern of speaker B when reading is faster, less careful and her utterances are correctly classified only by a small number of listeners. Thus, it may be concluded that listeners need contrasting patterns to identify utterances as spontaneous or read.

3.2 Results for the whole corpus

Assuming a Gaussian distribution, the speakers A and B would be situated in the extreme positions on both ends of the scale. To look for speaker-independent results and in order to find characteristics in the different speech types the examination of all speakers is indispensable. Therefore, the results for all four dialogue types will be illustrated.

3.2.1 Timing structure

Table 7 illustrates the results for articulation rate. As pauses and interrupting parts of the utterances were excluded for the measurement, articulation rate was expected to yield promising results. Comparing the values for the average as well as for median or standard deviation it is evident that there are no remarkable differences between the 4 styles, i.e. in all dialogue types almost 15 sounds/s are produced. From this it must be concluded that the results are speaker-specific.

It is not very surprising that the results for speech rate (table 8) are unsatisfactory since the examination included also non-fluent utterance parts which vary widely in frequency and length for the different speakers.

In order to shed more light on timing structure it is indispensable to have a look at the distribution of the "interrupting parts", i.e. pauses and sound-lengthenings.

Table 7: *Articulation rate* (number of sounds/s for all speakers; style-related)

Dialogue types	average	standard deviation	median	minimum	maximum	number of speakers
dialogue type 1	14.59	1.05	14.55	13.02	16.87	14
dialogue type 2	14.86	1.05	14.43	13.39	16.59	14
dialogue type 3	14.55	0.96	14.46	12.75	16.13	14
dialogue type 4	14.84	1.49	14.92	12.62	16.72	14

Table 8: *Speech rate* (number of labels/s for all speakers; style-related)

Dialogue types	average	standard deviation	median	minimum	maximum	number of speakers
dialogue type 1	11.52	1.01	11.40	10.24	13.98	14
dialogue type 2	11.27	1.58	11.37	8.96	13.92	14
dialogue type 3	11.43	1.58	11.62	7.23	13.42	14
dialogue type 4	12.53	1.49	12.35	9.34	14.63	14

3.2.1.1 Distribution of pauses

Pauses were divided into breathing and silent and filled pauses (hesitations). The examination revealed that global results for the speech corpus can be drawn, but they must not obscure the fact that the distribution of pauses is speaker-specific.

Comparing the three pause types, a hierarchy can be observed: In all dialogue types, breathing pauses are most frequent, followed by silent pauses, followed by filled pauses (table 9).

Table 9: Length of pauses related to the total time of speaking for all speakers; style-related

Pause type average in %	dialogue type 1 ("casual")	dialogue type 2 ("careful")	dialogue type 3 ("careful")	dialogue type 4 (read)
breathing	5.08	7.11	6.67	6.14
silent pauses	6.46	4.87	4.95	2.21
filled pauses	2.30	3.25	2.52	1.15

Table 10: Pauses in non-interrupting position related to the total amount of each pause type; style-related

Pause type average in %	dialogue type 1 ("casual")	dialogue type 2 ("careful")	dialogue type 3 ("careful")	dialogue type 4 (read)
breathing	68.47	79.06	80.75	84.88
silent pauses	59.52	58.65	69.79	86.31
filled pauses	32.56	47.54	59.68	75.97

Breathing occurs in all dialogue types mostly *between* fluent utterance parts, i.e. non-interrupting (cf. table 10). This is found increasingly the less spontaneous the communication situation. It is clear that the structural planning in read speech is much lower than in spontaneous speech and, thus, in read speech speakers obviously breathe *before* a new phrase is uttered (85% non-interrupting). In the casual speech type the amount of non-interrupting breathing pauses is lower (68%, cf. table 10) as the planning may require more attention.

The examination of **silent pauses** yields interesting results as we find a strong decrease of occurence from casual over the two careful dialogue types towards read speech (table 9). The lack of silent pauses in read speech may be explained by reading skills of the speakers. Normally, reading should also contain a certain amount of silent pauses. In non-professional reading we often find a lack of pauses. Concerning the position of silent pauses, they are largely non-interrupting in read speech (86%) which is due to the same reason as for breathing in this case. In the casual dialogue type a smaller amount of silent pauses is non-interrupting (60%). Maybe the structural planning of the utterances provokes the pauses also *within* phrases.

The least frequent pause type in all dialogue types are **filled pauses** (table 9). Comparing the dialogue types, they are most frequent in the careful speech style (dialogue type 2). This may be due to the subject of the dialogue which was new and uncommon for the speakers. Filled pauses occur in all speech styles apart from casual speech largely non-interrupting. The results reveal that the amount of interrupting hesitations is higher than for silent and breathing pauses. For casual speech it is interesting to observe that filled pauses occur mostly *within* phrases (only 33% non-interrupting, cf. table 10). This may be due to the fact that the communication situation was absolutely spontaneous without any tasks for the speakers, such that a role change was possible as often as the speakers intended to and, therefore, phrasing was interrupted.

3.2.1.2 Sound-lengthenings

Sound-lengthenings are found to occur to some extent in all speech styles and the amount is similar: about 2-3% of sounds are lengthened compared to the total number of sounds produced in a single dialogue type.

The average in the four dialogue types is 1.9% (1) - 3.0% (2) - 2.4% (3) - 2.2% (4). It is difficult to give a reliable interpretation. An analysis of the results for the single speakers reveals that the frequency of sound-lengthenings does not vary as strongly as for other parameters. As the lowest value in the total corpus is found in casual speech it may be concluded that the most natural communication situation provokes the least lengthenings. Generally, the occurence of lengthenings seems to be largely independent from speech styles and rather due to speech rate and fluency of an individual speaker. With regard to these very close results in all dialogue types it is necessary to have a look at the position of occurence.

Generally, lengthenings occur more and more before syntactic boundaries the less spontaneous the speech style. Whereas in casual speech the amount of non-interrupting lengthenings is about 45%, in read speech non-interrupting lengthenings take the most part (75%). In casual speech the reason for the high amount of interrupting lengthenings may be the structural planning of the utterances. However, it must not be neglected that 22% of lengthenings occur utterance-finally. It is interesting to note that in dialogue type 3, where the recording conditions had been strongly controlled, the amount of utterance-final lengthenings decreased (only 8%) compared to the dialogue types 1 and 2. The task for the speakers provoked different structural planning and led to longer utterances than in the other dialogues. It might be for the same reason that phrase-final utterances followed by a pause take the most part in all dialogue types. It is evident that the lack of structural planning in read speech yields an only small amount of interrupting lengthenings.

The reasons for sound-lengthenings vary with respect to articulation rate, the subject of the dialogue, the relation between the speakers, speech habits and the communication situation. The results for sound-lengthenings once more reveal that each communication situation has to be seen on its own.

3.2.2 Speech production

Reductions were measured as deletions, substitutions and insertions of sounds. The results for all three reduction types will be discussed.

Insertions cannot be classified as reductions, but they were measured as modifications of speech production in the four dialogue types. The examination revealed that in all dialogue types to some extent sounds are inserted. It can be interpreted that insertions occur rather due to the phoneme context and to the individual speaker than due to the communication situation.

In table 11 the average of deletions and substitutions related to the total number of sounds produced by all speakers is illustrated. Both reductions occur for the most part in casual speech and least in read speech. At first sight this result looks very clear, so that casual speech is to be judged as less carefully articulated than read speech. As we have learned from the listening experiment, articulation was not the only feature for the listeners to classify an utterance as spontaneous or read. Moreover, timing structure yielded different results for the single speakers. Furthermore, reading skills and individual speaking styles must not be neglected as influencing features on speech production.

Table 11: *Deletions* and *substitutions* related to the total number of sounds produced by all speakers

Reductions (average in %)	dialogue type 1 ("casual")	dialogue type 2 ("careful")	dialogue type 3 ("careful")	dialogue type 4 (read)
deletions	14.46	8.18	8.34	6.87
substitutions	10.70	6.75	6.95	6.28

Comparing the four dialogue types it can be concluded that to some extent the tasks of speaking were responsible for the occurence of reductions. It is important to note that the amount of deletions, and also substitutions, is definitely higher in casual speech than in the careful spontaneous speech styles and, of course, in read speech as well. This may be due to the very natural communication in the first dialogue type that in some cases led the speakers to a rather informal situation.

The close results in the dialogue types 2 and 3 reflect the similar tasks of the dialogue subject in both situations. For read speech it can be interpreted that the concentration on the printed text and the engagement in a rather careful reading yielded less reductions.

To explain reductions in more detail it is necessary to examine their distribution according to sound categories. Concerning deletions, in all dialogue types mostly plosives are deleted which often concern word-final plosives. After plosives, the next most frequently deleted sounds are vowels. This is due to the fact that syllables are contracted such that the vowel is no longer present. This manner of reduction often occured across word-boundaries, e.g. in the sequences "ja aber" (yes, but) or "aber ich" (but I). Especially in less formal communication situations (casual style, but also the dialogue types 2 and 3) vowels often are deleted when they occur as inflectional endings of verbs, e.g. "denk' ich" instead of "denke ich" (I think), "wär's" instead of "wäre es" (it would be). Substituted sounds are for the most part vowels. This can be explained by taking into account that vowels are influenced by the phoneme context and vowel context, respectively, i.e. in most cases vowels are assimilations of the surrounding syllables.

The examination of speaker-specific results in speech production reveals that deletions and substitutions are very speaker-dependent although the global result for the total corpus gives more evidence than did the other parameters that were measured.

4. Conclusions and discussion

The experiment revealed interesting results concerning the variability of speech production in different speech styles. A particular claim of the study was to see whether there are crucial differences between read speech and speech uttered spontaneously. The listening experiment yielded contrasting results for two selected speakers. Of course, apart from the selection of

utterances and the acoustical conditions, these results might be influenced by reading skills. If reading skills are to be defined as the ability of perfectly transferring visual print patterns into acoustic patterns, then the utterances of speaker A should be considered as a perfect reading. The utterances of speaker B, however, are much more fluent and more natural-sounding than those of speaker A, although (or because) they contain more reductions. As the speech patterns of the two speakers are the most contrasting in the total corpus, the range of variability independent from the dialogue types is extensive. Thus, generally, read and spontaneous utterances are difficult to distinguish.

The listening experiment also reveals the way listeners expect speech to be, i.e. read speech as rather slow, carefully articulated and very intelligible, and spontaneous speech as rather slurred and not quite well articulated and, moreover, containing pauses and interruptions. The conclusion drawn from the listening experiment is, thus, that listeners are able to identify spontaneous and read utterances only if the differences are evident.

The answer to the question what about the range of variablity within a style is to be seen as speaker-specific. There are speakers whose speech patterns are even contrasting within the same style.

Concerning speech production, the patterns both in articulation and timing structure changing from one style to another are not corresponding for all speakers. Whereas articulation rate is very speaker-dependent, speech production reveals a tendency of "reduction patterns": more reductions occur in spontaneous speech than in read speech, but, keeping in mind reading skills and the claim to articulate carefully when reading.

The question whether there are crucial differences in read and spontaneous speech is to be answered as follows: The distinction between spontaneous and read is superficial, and is not a classification of articulatory variability. Furthermore, speech utterances are adapted to a given particular communication situation.

Therefore, speech should be seen as a spectrum of styles depending on situational facts and on speaker-specific facts, i.e. there are as many speech styles as there are communication situations.

Acknowledgement

This research was partly supported by the German Federal Ministry of Education, Science, Research and Technology (BMBF).

References

Beckman, M.E. (1995): A typology of spontaneous speech. In: Proc. ATR International Workshop on computational modeling of prosody for spontaneous speech processing, Kyoto, 2-23 - 2-34.

Eskénazi, M. (1993): Trends in Speaking Styles Research. In: Proc. ESCA Eurospeech, Vol. 1, 501 - 508.

Stock, D. (1992): P-TRA - Eine Programmiersprache zur phonetischen Transkription. In: Beiträge zur angewandten und experimentellen Phonetik. Steiner Verlag, Stuttgart.

Some observations on *'ein'* and *'einen'*

Bernd Pompino-Marschall and Peter M. Janker
Zentrum für Allgemeine Sprachwissenschaft, Belin

Introduction

The paper addresses the question of what distinguishes the two word forms *'ein'* and *'einen'*, i.e. the German indefinite article nominative singular (mask., neutr.) vs. accusative singular (mask.) when produced in connected speech.

Generally, it is assumed that [ən] endings in German by [ə] deletion reduce to syllabic [n̩] in casual pronunciation. This for example seems to be litterally true for German content word forms ending in [nən]. The material analysed in preparation of the planned new edition of the WdA in about 80% of the cases showed [ə] deletion and more than 50% of these reduced forms showed clear indications of two separable [n] sounds, i.e. differences in the energy or f_0 contour (Siegrun Lemke, personal communication). For the function word contrast between *'ein'* and *'einen'* however, inspection of the Kiel corpus of read speech in almost no cases of reduced *'einen'* revealed reflexes of two [n] sounds in the energy or f_0 contour even in cases where two [n] segments were labelled.[1] The distinction of these word form tokens from *'ein'* therefore seems to rely on the different length of the [n] sounds, i.e. [aɪn:] vs. [aɪn] (as against assumed [aɪnn̩] vs. [aɪn]).

1 The perception of *'ein'* and *'einen'*

In order to test this hypothesis that hearers are capable of differentiate between both word forms cued by [n] duration only, we constructed a couple of listening experiments with manipulated naturally produced acoustic speech material. A prototypical token of reduced 'einen' was cut from the utterance (RTDS046) with the help of the Signalyze 3.12 software for Apple Macintosh. It consisted of a glottalized [a] segment of 68 ms, a modal diphthongal segment of 65 ms and a [n] segment of 49 ms duration (cf. figure 1).

Procedure

The duration of the word final [n] was modified the following way: (1) two step shortening of 10 ms each by cutting the fifth and sixth or the fourth to seventh pitch period counted from the end of the acoustic signal and (2) two step lengthening by doubling the pitch periods five and six or four to seven. Since it was expected that the duration based distinction is dependent on speech rate, the duration of the diphthongal segment was also manipulated in equal steps of 10 ms by cutting/doubling two or four pitch periods. The pitch periods for this manipulation were chosen in a way to keep the formant transition of the diphthong intact (cf. figure 1). This resulted in 25 stimuli (5 [n] durations * 5 [aɪ] durations: -20, -10ms, original, +10, +20 ms each) that were presented with and without the initial glottalisation five times in quasi randomized order (resulting in 250 items) to the subjects. The subjects listened to the stimuli via headphones at a comfortable listening level in a quiet room. They marked their identification responses (*'ein'* or *'einen'*) on prepared answer sheets.

---

[1] Here, some slight spectral changes at the supposed segment boundary may be detectable.

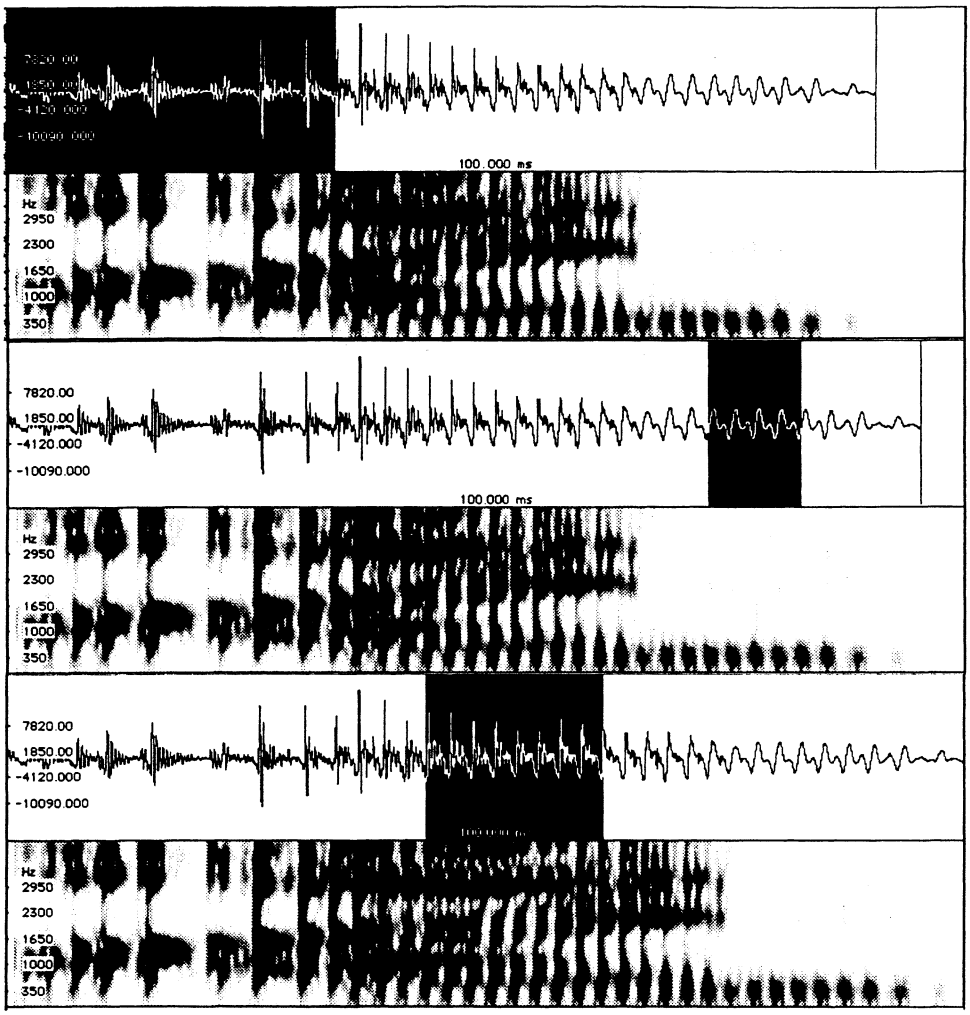


Fig. 1: Oscillograms and sonagrams of selected stimuli: The original signal with the initial glottalisation marked (top); first step of [n] lengthening with the two doubled pitch periods marked (mid); second step of [aɪ] lengthening with the four duplicated pitch periods marked (bottom).

Results

The results of this listening test averaged over 17 subjects are depicted in figure 2. The raw data was subjected to an analysis of variance with the factors [n] duration, [aɪ] duration and glottalisation and the number of *'einen'* responses as the dependent variable.

Analyses of variance revealed a highly significant ($p << .001$) effect of [n] duration, presence of glottalization as well as a significant ($p < .01$) interaction of both effects and a marginal effect ($p < .05$) of vowel length (distinguishing between vowel length 2 and 5; cf. fig. 1). Glottalized items were generally more often identified as *'einen'*. For the glottalized as well as for the non-glottalized items only a highly significant ($p << .001$) effect of [n] duration remained: The longer the duration of the nasal segment the more *'einen'* responses. For post hoc Scheffe comparisons of pairs for glottalized items only pair 1/2 and 4/5 and for non-glottalized items additionally pair 2/3 failed to reach significance.

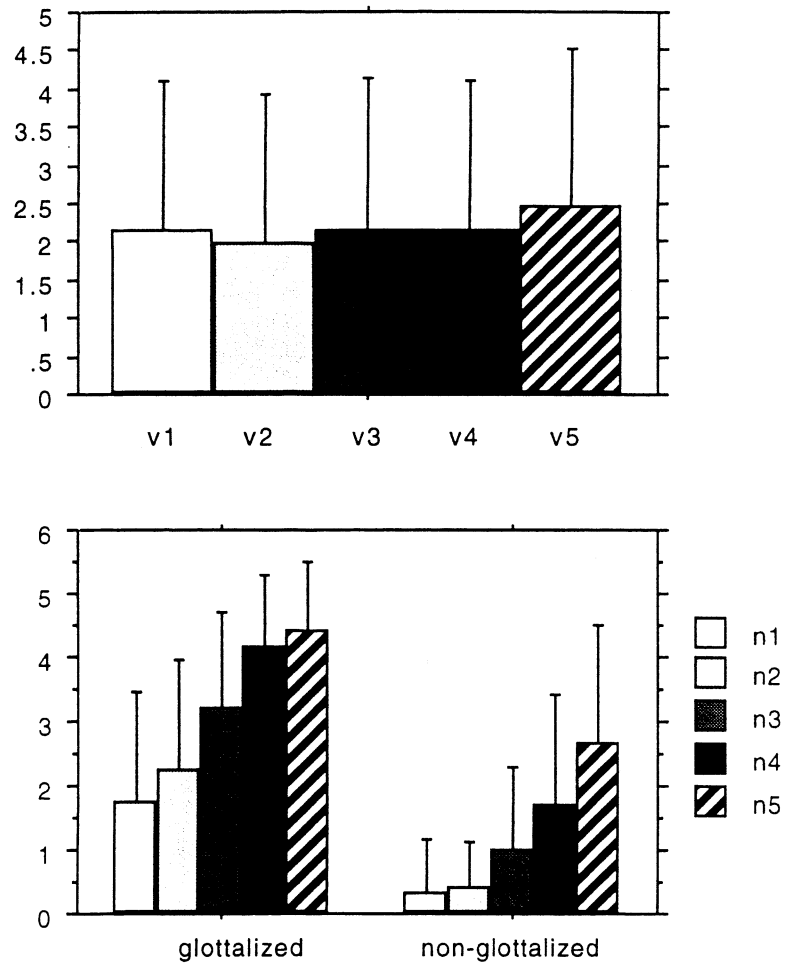


Fig. 2: Results of the listening experiment: Mean *'einen'* responses to stimuli with diphthongal segments of different length (top); interaction of [n] duration and presence/absence of glottalisation (bottom; error bars represent one s.d.).

For the perception of socalled syllabic [n] in our material the duration of the nasal segment - independent of the duration of the modally voiced vocalic portion but reinforced by the presence of glottalization (i.e. a longer vocalic portion) and therefore independent of speech rate - seems to be the only reliable segmental cue. We will take up the effect of glottalization again in the general discussion.

2 The production of *'ein'* and *'einen'*

Recording procedure

Figure 3 depicts the general experimental setup. Tongue movements were monitored by means of electromagnetic articulography (AG100 Carstens Medizinelektronik, Göttingen, Germany). This method involves the use of three transmitter coils (mounted on a helmet) to generate an alternating magnetic field at three different frequencies. The field strength detected by sensor coils mounted on the articulators is roughly inversely proportional to the cube of the distance between sensor and transmitter (see Perkell et al. 1992, 1993; Schönle 1988 for background to

electromagnetic transduction systems). The raw distance signals are then converted by software to x-y coordinates in the midsagittal plane. In order to guarantee the quality of the articulatory data, additional procedures were implemented allowing more accurate calibration and better detection of unreliable data (see Hoole 1993 for details).

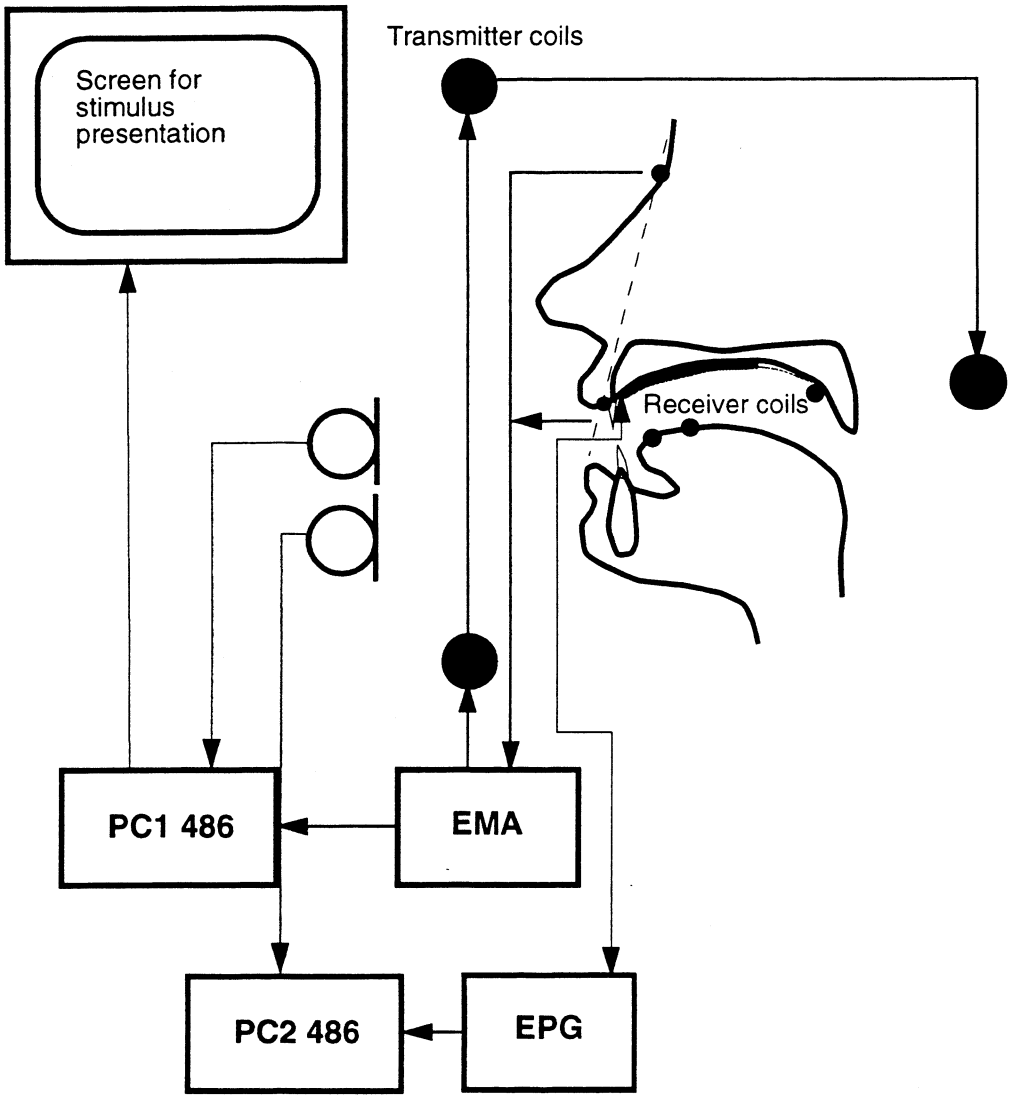


Fig. 3: Experimental setup and placement of receiver coils (bottom right: front view of the subject with tongue streched out to demonstrate the placement of the receiver coils).

Details of the sensor positions are as follows: Two transducers were mounted on the midline of the tongue at about 1 and 5 cm from the tongue tip (henceforth TB - tongue blade - and TD - tongue dorsum coil, respectively). The third coil was mounted on a strip of elastic foil glued to the back of the artificial EPG palate touching against the back part of the velum when the palate is inserted (henceforth V; cf. figure 4). Two reference coils were attached to the upper incisors and the bridge of the nose to correct for head movements.

The modified recording software (Hoole 1993) stored the movement data of the five receiver coils (recorded at 400 Hz) together with the information of the instantaneous tilt and the synchronous audio signal (16 bit, 16 kHz) in compressed form on a PC.

Besides the articulatory data at the end of the test session a tracing of the hard palate of the subject was made by using a sensor attached to the finger of one of the investigators.

The raw data were preprocessed to (1) correct for the remaining measurement error[2], (2) rotate to the vertical axis defined by the positions of the coils at the bridge of the nose and the upper incisors, (3) decompress the audio file, and (4) splitting the tilt data from the position data.

Tongue palate contacts were measured by means of the Reading EPG3 system with an artifical palate with 62 electrodes (seven rows of eight plus front row of six electrodes) every ten ms parallel to the audio recording with another PC (cf. figure 3 & 4).

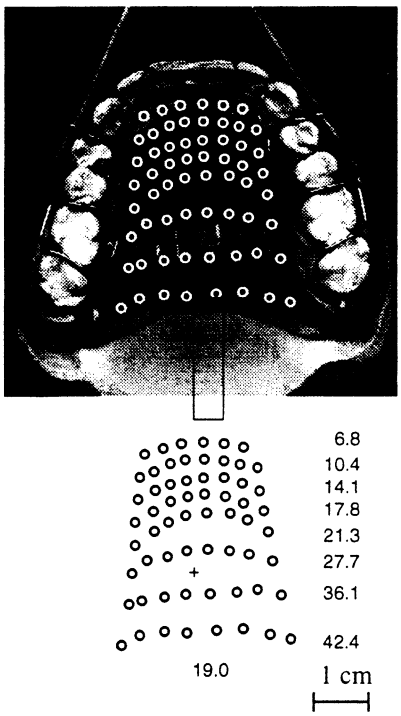


Fig. 4: EPG palate (above) and scheme of electrode placement (below) for subject JDR (the rectangle marks the strip of foil glued to the palate to carry the velar EMA coil; mean distances [in mm] of electrode rows from the inner edge of the upper incisors is given right, the higest point from bite plane [distance in mm given below] is marked by a cross).

Material

The male native German subject (JDR) read parallel constructed sentences with 'ein' and 'einen' in randomized order five times each in three different recording blocks. First in his normal pronunciation, the second time more carefully and then again more quickly and casually. The sentences had the form of

'Es fuhr ein Audi nach Augsburg' ('An Audi was going to Augsburg')
'Er fuhr einen Audi nach Augsburg' (He drove an Audi to Augsburg)

(parallel sentences contained *'Kombi'* ('utility car') / *'Cottbus'*, *'Traktor'* ('tractor') / *'Trabach'*, *'Volvo'* / *'Wolfsburg'*, *'Mazda'* / *'Monza'*).

[2] By using the computed error during calibration.

Analysis procedures

Durational measurements in the acoustical signal were conducted with the Signalyze software for Apple Macintosh. In the acoustic signal the following durations/time points were determined manually under auditory and visual (especially sonagraphic) feedback: the beginning of the initial glottalisation, the beginning of the diphthong, the beginning and the end of the first [n] segment as well as (when applicable) the beginning of the [ə] segement and/or the beginning and the end of the second [n] segment.

The EPG data was analyzed (1) with respect to the duration of linguopalatal contact during the production of the nasals and (2) with respect to the position of the centre of gravity of the area of linguopalatal contact averaged over all frames that show at least one electrode row of total closure. As in Gibbon et al. (1993) the centre of gravity was computed over the contacts of the central four midsagittal electrodes of the anterior fo rows of electrodes. In case of [ə] elision in items of *'einen'* the centre of gravity was calculated separately for the first and the second half of the total contact duration. For cases of assimilatory changing of the nasal place of articulation (in items of *'Kombi'* / *'Cottbus'*) also the duration of closure overlap at the alveolar and the velar place of articulation was measured.

The EMA data was analyzed with respect to the alveolar closing/opening behaviour of the TB coil and the velar lowering/raising behaviour of the V coil. Minima within the tangential velecity function were used as starting points for gestural analysis. On- and offsets of movement were defined as 20% points within the total flesh point displacement function. Besides the maximal velocity the duration as well as the fleshpoint position at the beginning and the end of the gesture and the intergestural timing (e.g. velar hold - the interval between the end of velar lowering and the beginning of velar raising) were determined.

The data were subjected to analyses of variance with speaking style (careful, normal, fast), produced word form (*'ein'*, *'einen'*) and following consonant (zero, labial, dental, alveolar, velar) as independent variables.

2 Results

The results of the durational measurements on the audio signal are reported in table 1.

Table 1: Acoustical segment durations

word	style	~	aɪ	n	ə	n
ein	careful	23.4 (15.75)	55.9 (21.87)	83.0 (21.78)		
	normal	34.8 (16.72)	83.5 (15.61)	98.6 (29.66)		
	fast	.4 (2.10)	51.3 (11.49)	58.2 (14.43)		
einen	careful	24.6 (15.59)	47.5 (15.25)	62.2 (70.63)	73.8 (19.36)	79.1 (25.03)
	normal	27.4 (13.98)	62.0 (15.48)	248.1 (41.08)	•	•
	fast	1.0 (3.56)	61.6 (11.89)	107.0 (24.51)	•	•

Analyses of variance revealed a highly significant ($p << .001$) effect of speaking style on the duration of the word-initial glottalization due to the fact that it is nearly totally absent (in 94% of the cases) in the fast productions of our speaker. The duration of the modal diphthong as well as the vocalic part as a whole showed a higly significant ($p << .001$) effect of speaking style, a

significant ($p < .05$) effect of word form and also a higly significant ($p << .001$) interaction of both effects. These vocalic parts are significantly ($p < .01$) shorter (-21.6/-29.0 ms) in the normal but also significantly ($p < .01$) longer (10.4/11.0 ms) in the fast productions of the word form *'einen'* than in *'ein'*. The (first) nasal segment showed highly significant ($p << .001$) influences of speaking style for both word forms and for *'ein'* also of consonantal context ($p << .001$) as well as an interaction of both effects ($p < .01$). Simple effects show up as different significant differences (none in alveolar context) within the general ranking 'fast < careful < normal' and significant differences in the ranking between alveolar and labiodental (mean difference 43.2 ms).

A preliminary analysis of the EPG and EMA data showed results that in some cases seem to contradict the acoustical measurements. So, for example, in seven cases (i.e. 28%) of the normal *'einen'* productions there was no perfect alveolar closing contact resulting in different segmental durations when measured by EPG. For these items, on the other hand, the EMA data didn't show significant differences in the amount of vertical movement of the TB coil.

The EMA analysis of the *'einen'* utterances revealed that despite an always present slight tongue tip lowering of about 4 mm (in contrast to the elevation of 11.6 mm for producing the alveolar closure of the [n]) for the [ə] production in careful style, the velum does not show corresponding closing movements but remains open during the vowel production. This velar lowering is, on the other hand, with 1.9 mm significantly less ($p << .001$) than for the normal and fast productions (3.8 mm) and starts on the average 29.3 ms later than the tongue blade movement in significant contrast ($p << .001$) to the normal and fast productions, where the velar gesture precedes the tongue blade movement by about 96.3 ms. The tongue tip lowering for [ə] was only once observed in normal speaking style, never at fast speech rate.

As to be expected, the variation of the consonantal context affected the position of the alveolar contact for the nasal: The position of the alveolar closure as determined by the center of gravity of the EPG contact pattern showed significant influences of the speaking style, the consonantal context as well as their interaction for the whole nasal segment in *'ein'* ($p << .001$; $p < .05$; $p < .05$) as well as the second (or the second half of the) nasal segment in *'einen'* ($p << .001$; $p < .01$; $p << .001$). Split by word form and speaking style the simple effects were as follows: The contact in the nasal segment of *'ein'* in the velar context is 0.18 rows more backwards than in the zero and alveolar context in careful pronunciations ($p < .01$), 0.30 and 0.32 rows more backwards than in all other contexts in normal and fast productions (both $p << .001$); for the careful *'einen'* utterances there was an only marginal ($p < .05$) effect (zero context 0.09 more backwards than alveolar context), but for the normal and fast productions the velar context again showed more backward contacts in the velar context (0.28 rows in contrast to all other contexts for the normal productions and 0.13 rows in contrast to the labial and labiodental context; both $p < .001$).

The nasal productions in the velar context also showed an assimilatory overlapping (of about 55 ms) of alveolar and velar contacts that is marginally significant ($p < .05$) dependent on speaking style: With 64 ms this overlap is 19 ms longer in fast speech rate than in normal productions.

4 General discussion

The experiments described above demonstrate the large variability of word form realizations in this quite simple example of *'ein'* vs. *'einen'*.

For the hearer, the most reliable cue for bisyllabic *'einen'* seems to be the pure length of the nasal segment. But there is also an effect of the presence of an initial glottalization.[3] These more frequent 'einen' responses to glottalized items may be due to the fact that this glottalization is perceived as an event of its own, not belonging to the following syllable, i.e. resulting in a bisyllabic percept. To cite an example of a natural realization of *'einen'* already described in Pompino-Marschall (1996) there seems to be the possibility of signalling the bisyllabicity of *'einen'* not by a syllabic [n̩] but - anticipated - as a bisyllabic vocalic segment: The original utterance shown in figure 5 is unambiguously perceived as *"Gibt es einen Zug ...?"* but when the steady-state [a] portion of 49 ms duration is cut - as shown in figure 5 (top) - the ungrammatical utterance * *"Gibt es ein Zug ...?"* is heard.

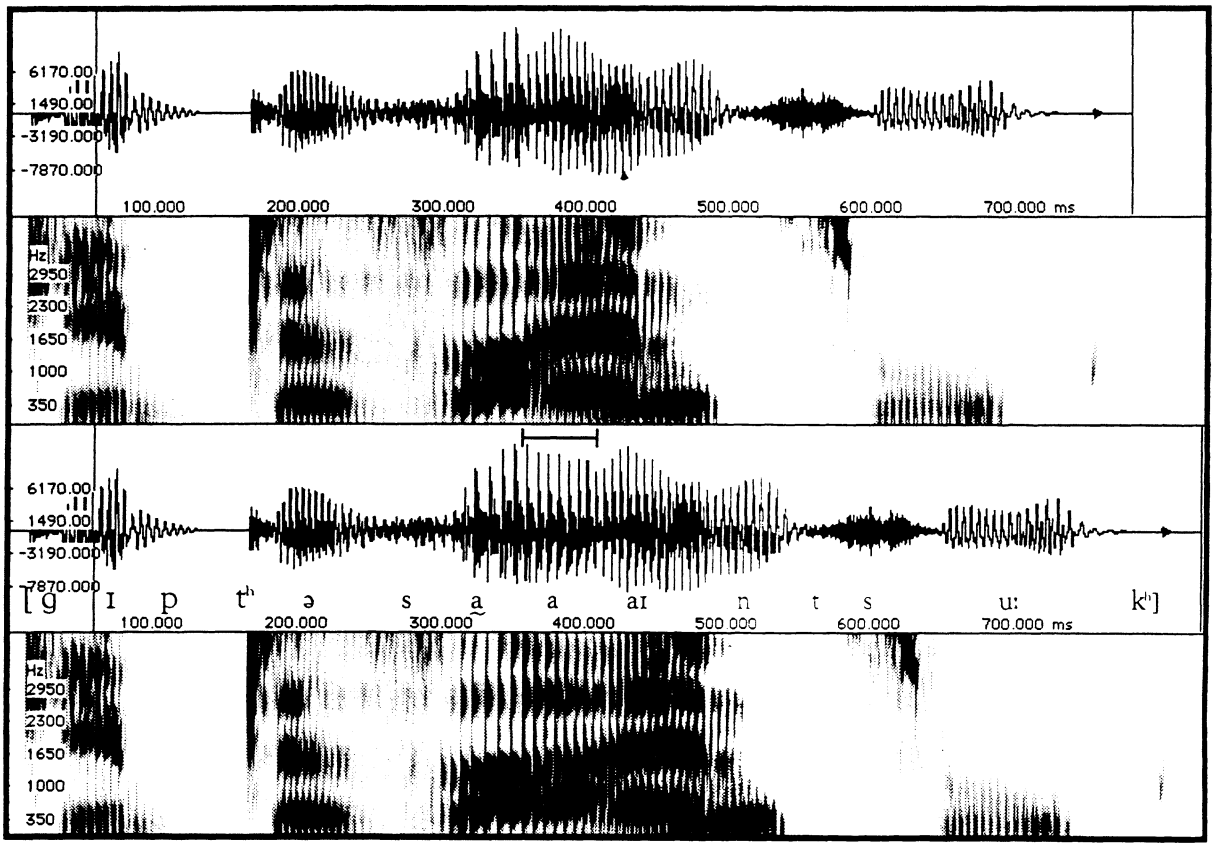


Fig. : Audio signals and sonagrams of the manipulated (top) and the original (bottom) utterance "Gibt es einen Zug ...?" (HPTS063): In the manipulated utterance the marked [a] segment of 49 ms was cut.

In order to test this possibility further, some more test tapes were prepared where the preceding word *'noch'* of the original utterance was pasted before the manipulated items of *'einen'* of the listening tests described above. This procedure yielded stimuli in which the glottalization is perceived as an integral part of the following syllable. To compensate for a perceptually resulting speech rate accelleration in the stimuli with deleted glottalizations in these items also a silent

[3] N.B. in opposite direction of a compensatory shortening of the vocalic part of the utterance.

interval of the duration of the glottalized segment was inserted. To test the influence of the diphthong - besides its duration - on the perceived dichotomy another test with synthetic material was constructed where the amount of the F1/F2-transition from [a] to [ɪ] was varied systematically.

The articulatory investigation, on the other hand, showed that it is by no means a simple segmental process that underlies the different pronunciation variants of *'einen'*. There is not only a simple deletion or reduction of single gestures but also a complex restructuring of the inter-articulator coordination in timing as well as in amount.

References

Gibbon, F.; Hardcastle, W. & Nicolaidis, K. (1993), Temporal and spatial aspects of lingual coarticulation in /kl/ sequences: A cross-linguistic investigation. Language and Speech 36, 261-277.

Hoole, P. (1993), Instrumentelle Untersuchungen in der artikulatorischen Phonetik: Überlegungen zu ihrem Stellenwert als Grundlage für Entwicklung und Einsatz eines Systems zur Analyse der räumlichen und zeitlichen Strukturierung von Sprechbewegungen. München [phil.Diss.].

IPDS (1995), CD-ROM #1: The Kiel Corpus of Read Speech, Vol. 1. Kiel.

Kohler, K.J. (1994), Lexica of the Kiel PHONDAT Corpus Read Speech, Kiel [= Arbeitsberichte Institut für Phonetik und digitale Sprachverarbeitung Universität Kiel (AIPUK) 27 & 28].

Kohler, K.J. (1996), Articulatory reduction in German spontaneous speech. In: Proceedings of the 1st ESCA Tutorial and Research Workshop on Speech Production Modelling & 4th Speech Production Seminar. Autrans, 1-4.

Perkell, J.S.; Cohen, M.; Svirsky, M.A.; Matthies, M.L.; Garabieta, I. & Jackson, M.T.T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. Journal of the Acoustical Society of America 92, 3078-3096.

Perkell, J.S.; Svirsky, M.A.; Matthies, M.L.& Manzella, J. (1993), On the use of electromagnetic midsagittal articulometer (EMMA) systems. In: Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM) 31, 29-42.

Pompino-Marschall, B. (1996), Articulatory reduction in fluent speech: A pilot study on syllabic [n] in Standard German. ZAS Papers in Linguistics (ZASPIL) 7, 151-162.

Schönle, P. (1988), Elektromagnetische Artikulographie. Berlin.

Aknowledgement

Work supported by German Research Concil (DFG) Grant Po 334/2 to the first author.