Evidence for the p-center syllable-nucleus-onset correspondence hypothesis

Peter M. Janker
Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung, Jägerstraße 10/11, D-10117 Berlin
janker@fas.ag-berlin.mpg.de

ABSTRACT

The question of whether or not the vowel onset is the parameter responsible for the rhythmical alignment in timing speech utterances still remains unclear. The experiment described here supports the hypothesis of a congruence between the mental representation of the syllable-nucleus-onset and the moment of occurrence or so-called p-center. Utterances consisting of phonologically identical but physically distinct syllable rhyme and different syllable initial consonance were presented in a synchronous tapping experiment. The results show that the vowel onset estimates the location of the tapping position and the differences between the stimuli as good as other models, suggesting a clear correspondence between p-center and vowel onset, thereby strongly supporting the hypothesis.

INTRODUCTION

The prosodic aspects of speech have gained more and more interest within recent years. Although the main interest was concerned with intonation patterns and tonal aspects the question of which parameters influence the rhythmic structure of utterances was not totally neglected. Since Terhardt/Schütte [1], Morton et al. [2] and Marcus [3] it is known that an acoustic (speech-) signal, like a syllable, is not perceived by its physical onset but somewhat later in time depending on its properties and/or physical make-up. Since then various investigations [4-13] have been undertaken and models proposed [3, 4, 10, 12, 14] to estimate this so-called 'Ereigniszeitpunkt', 'moment of occurrence or p-center' and to define the parameters influencing it. A single parameter could not be found, but it seems that the duration and/or physical make-up of the initial consonance of a syllable or the transition between consonance and vowel is of more importance than the syllable rhyme or coda.
But the question of which parameters influence the rhythmic structure of utterances was not only discussed in the light of the p-center/'Ereigniszeitpunkt' approach but also addressed by others, such as Allen [15], Rapp [16], Hollister [17] and Meyer [18]. Looking for physically measurable equivalents of the rhythm

beat, syllable beat, 'Silbenschlag' etc. these researchers asked subjects to tap[1] synchronously while listening to or speaking lyric / poetic utterances. They all agreed on that the 'beat' is somewhat near the end of the syllable initial consonance and the beginning of the following vowel as Meyer already stated in his 1898 paper "Über den Takt".
In some of my later work [9, 19-21] I argued for the hypothesis of a congruence of the p-center and the mental representation of the syllable-nucleus-onset and pointed out that the mismatch between tapping location and vowel onset found in these investigations might be due to not taking into account the phenomenon of anticipation occurring in tapping experiments (see 'Why anticipation occurs' below).

EXPERIMENTAL DESIGN

The experiment described here is undertaken to test the hypothesis that a congruence exists between the mental representation of the syllable-nucleus-onset and the p-center. In discussing results of former experiments I also suggested, that – besides neglecting the anticipation phenomenon – somewhat misleading results might occur due to the laboratory situation and the fact that some of the stimuli used in tapping and adjustment experiments might be somehow artificial lacking any compensational variations common in normal utterances, as they are, for the purpose of the experiment, varied in only one parameter and kept constant in the others. This experiment uses only segments of naturally spoken utterances as stimulus material.

Stimuli

The set of stimuli used here consists of eleven words with #CVC structure and two interjections [pstʰ, sːtʰ]. As the initial part of a syllable, the syllable head, has the main influence on the perception of the p-center position and the rest of the syllable has a lesser influence, the stimuli produced have different initial consonants [ʔ, p, f, k, h, kʰ, l, m, pʰ, ʁ, tʰ] and a constant syllable rhyme [astʰ], phonotactically the same but physically due to the natural generation different with all the compensational information of a natural utterance.
The words were well pronounced (explicitly demonstrated) in focus position within the frame sentence <Ich habe das Wort _____ gesagt.> (I said the word _____.)[2] and recorded in a soundproof studio using an Electro Voice 631B microphone and a DAT recorder. The interjections were produced as if the speaker wants to get someones attention [sːtʰ] or causes someone to be quiet [pstʰ]. The re-

1) For clarification: Alignment is the caused relation between two events (here sequences). Adjustment is an alignment achieved by changing the relation of two presented event strings, tapping an alignment achieved by performing an action (the tap) to a presented event string. The adjustment or tapping position (alignment position) is this alignment expressed in relation to an arbitrarily chosen origin. The location of the adjustment or tapping position is this measurement expressed in relation to other properties. Adjustment or tapping position are NOT the p-center position, rather they are induced by the perception of the p-center thus reflecting any change in the location of the p-center position.

2) N.b. that the position in the German phrase is not sentence final.

cording was transmitted to a Macintosh using Digidesign Audiomedia II (SP/DIF input), downsampled to 20 kHz and segmented using Signalyze on the Macintosh. The signal segments of interest were cut out of the sound stream and saved in a separate file for presentation. Figures 1 to 3 show as examples the

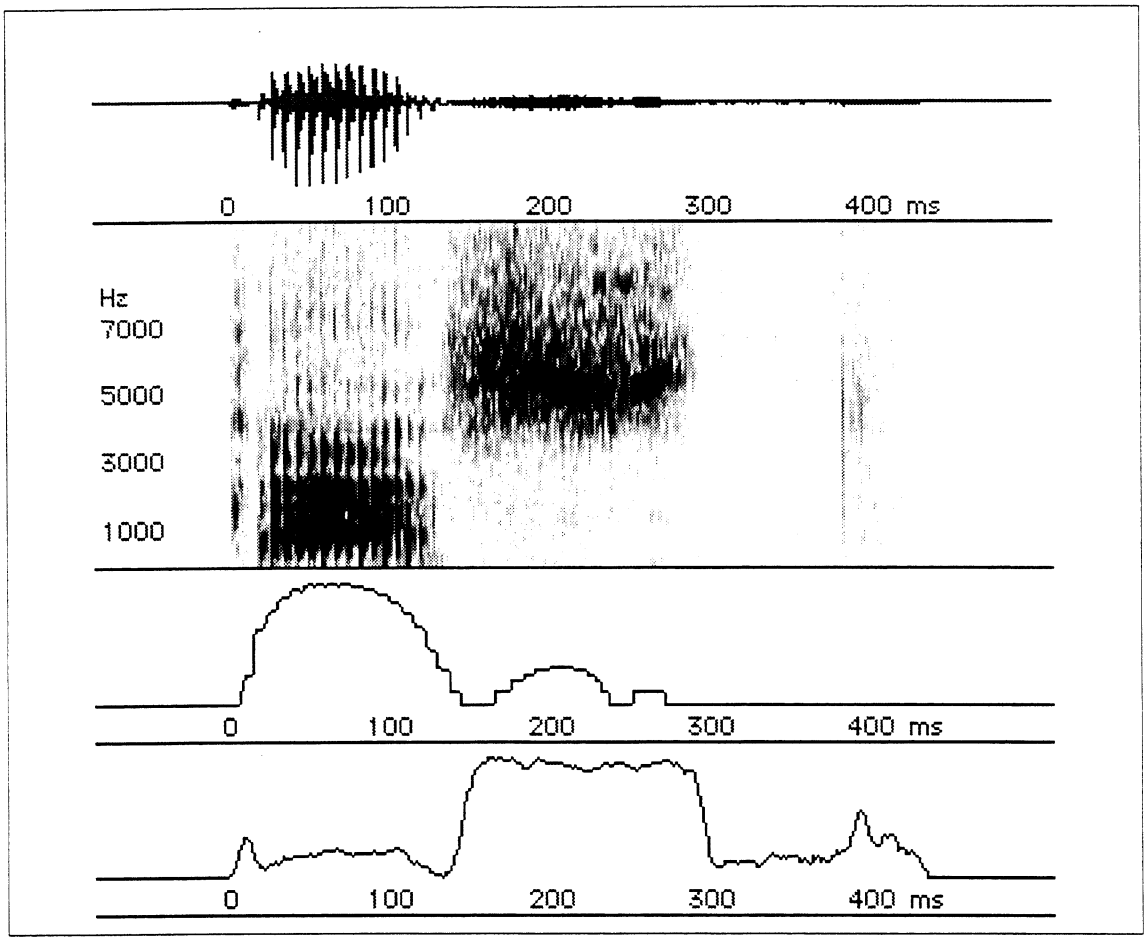

Figure 1: *<gast> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <gast> with an overall duration of 405 ms.*

shortest stimulus [kastʰ] with a duration of 405 ms, the longest stimulus [ʁastʰ] with an overall duration of 567 ms and the interjection [sːtʰ] with a duration of 416 ms. The figures show the amplitude as well as a wide band sonagram (10 kHz, pre-emphasis, 300 Hz window), the amplitude envelope (30 ms window) and the zero crossing (10 ms window) for the respective stimuli. An artificial 5 ms, 1 kHz tone burst perceived as a click signal was used as control stimulus.

Subjects

29 subjects (19 female, 10 male) took part in a series of experiments including this experiment. None of them had participated in experiments on rhythm perception before. To become familiar with the computer, the stimulus presentation program and the data acquisition procedure every subject had a 10 minute introduction using a click signal, the sound [pst] and the word <schwimmst> [ʃβɪmst] as stimulus material.

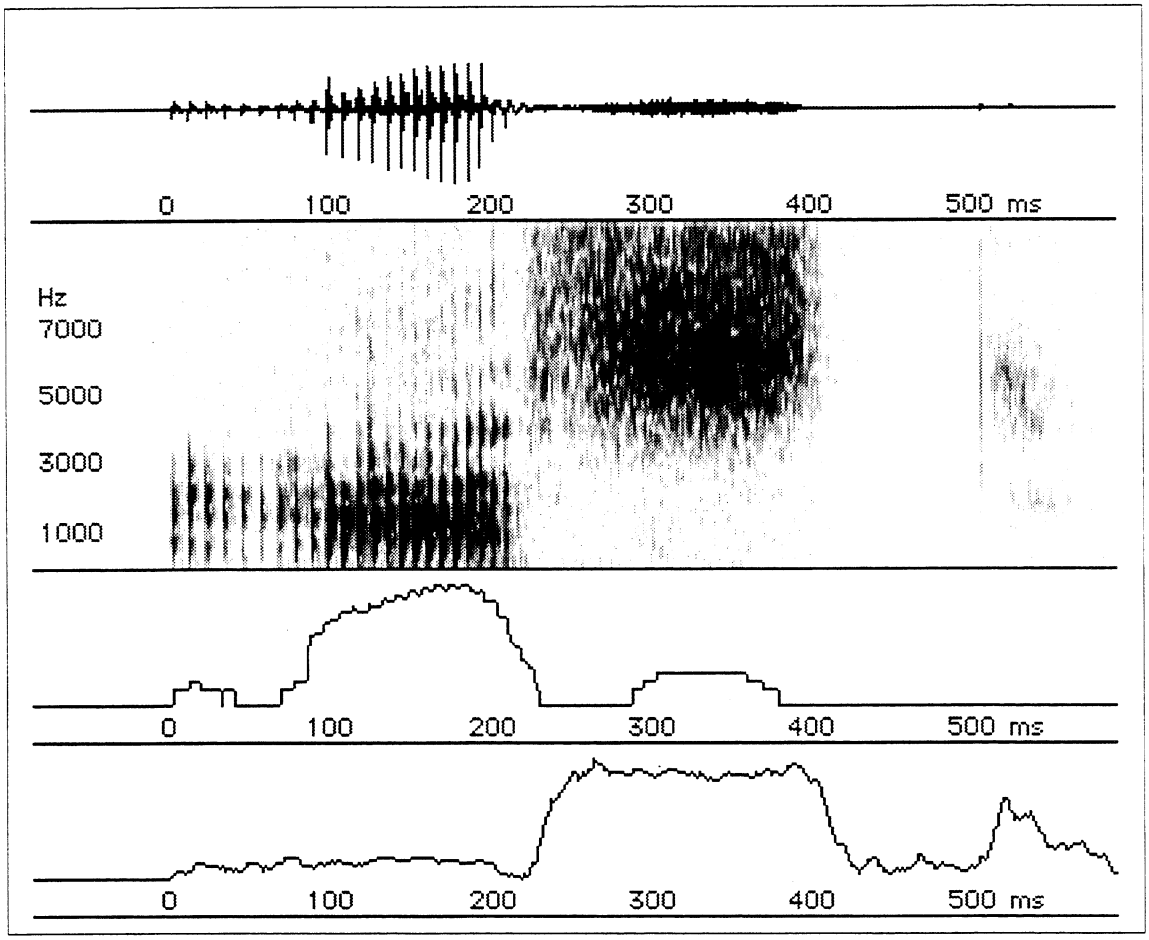

Figure 2: *<rast> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <rast> with an overall duration of 567 ms.*

Method

The stimuli were presented using Sennheiser HD 480 II headphones under computer control (Compac Deskpro 486/33, Data Translation DT2821SE DA-converter, Behringer PEQ 305 filter, Sony F535R) with 20 kHz sample rate and low-pass filtered at 6 kHz (24 dB/oct). The subjects had to perform a synchronisation task by tapping to sequences of binaurally presented stimuli. A sequence consisted of 15 repetitions of the same stimulus with an inter stimulus interval of 700 ms. The interval between sequences was 1400 ms. The stimulus sequences were randomized and grouped in blocks of 10. The subjects started the presentation of the next block by pressing the return key. A sequence was repeatedly presented as long as the subject did not start to tap. Each stimulus sequence was given four times with at least two different intermediate sequences. To register the taps a 5 x 10 cm capacitive sensory field was used. For analysis the taps to the first three and the last two presentations within a sequence were omitted (leaving 16,240 taps). Figure 4 gives a schematic description of the measurement procedure; a more detailed description of the experimental design and measurement procedure can be found in [6].

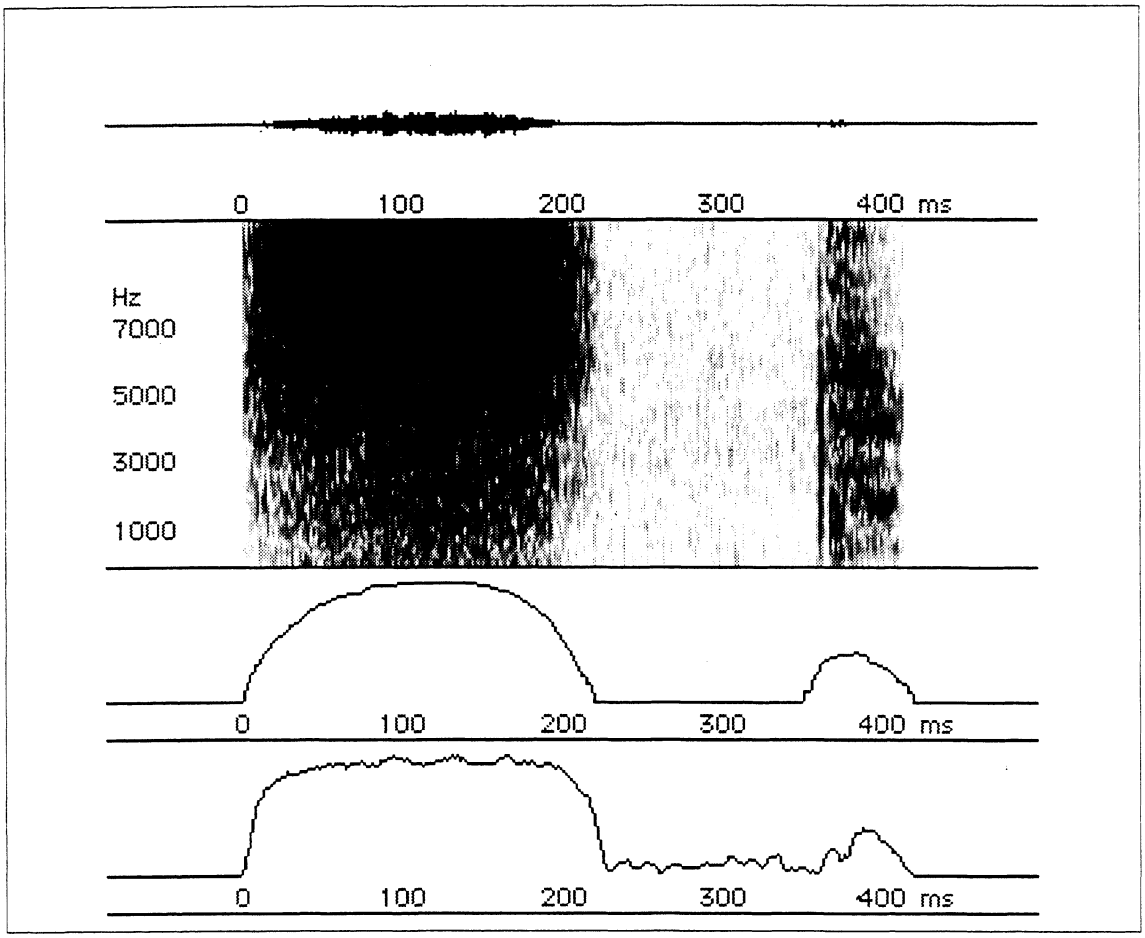

Figure 3: *<sst> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <sst> with an overall duration of 416 ms.*

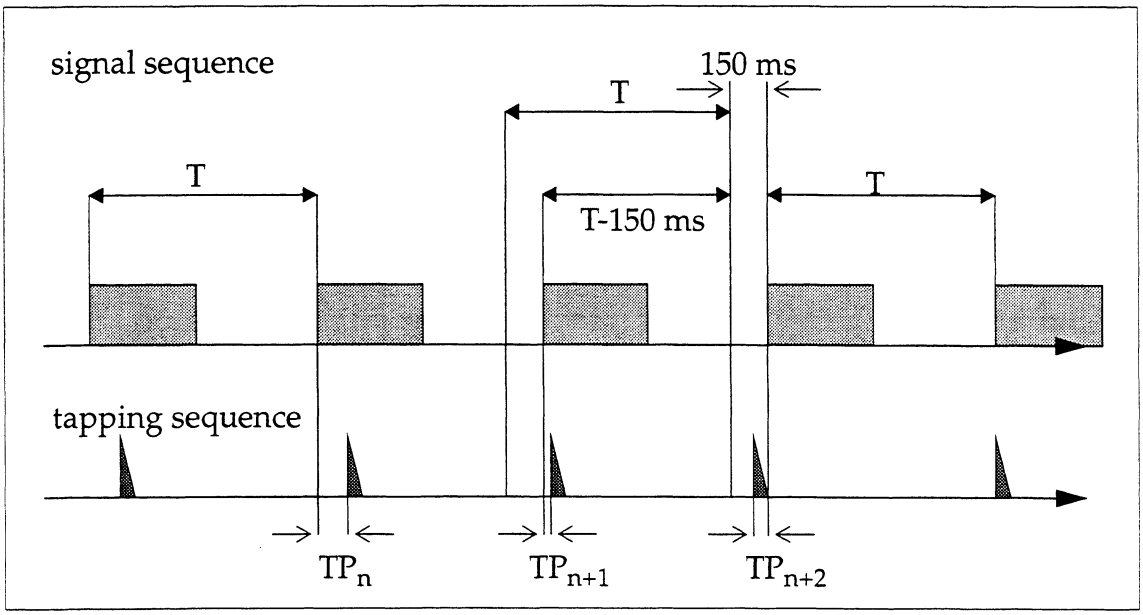

Figure 4: *Schematic description of tapping position measurement with tapping position TP_n in relation to the signal onset of the respective stimulus presentation (T = inter stimulus interval, TP_n = tapping position for the nth stimulus presentation, 150 ms offset for technical reasons)*

Results and Discussion

As the initial consonants of the stimuli presented had been different in duration and the distribution of spectral energy the repeated measurement design analysis gives a significant effect ($F = 99.79$, $p < .001$) for the factor stimulus as well as for the factor subject ($F = 18.05$, $p < .001$) with a significant interaction ($F = 1.49$, $p < .001$), as four subjects do not show a significant stimulus effect (post hoc Scheffé (.01)).
In tapping experiments the subjects have to perform three different tasks simultaneously. First to adopt a uniform rhythm given by an event string of identical stimulus repetitions, second to act to these events by tapping synchronously to the events and third to judge whether or not they succeeded in doing so. Therefore, a large difference in subject responses is typical for tapping experiments. To give an impression of the variability found in experiments like this, figure 5

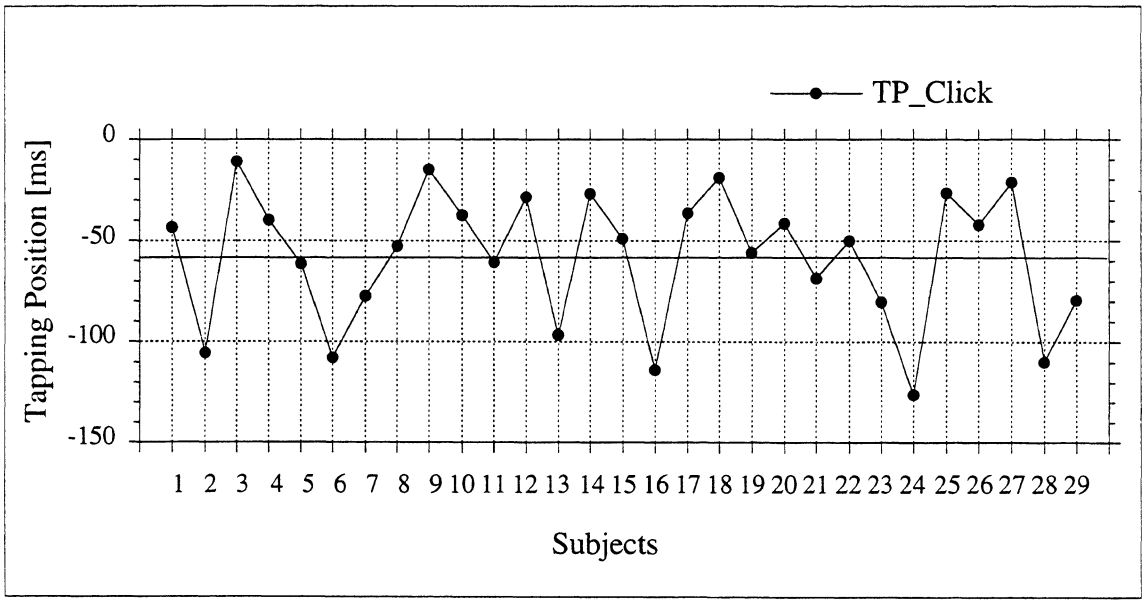

Figure 5: *Mean tapping positions TP for the control stimulus (click signal) for all subjects with indicated overall mean (solid line)*

shows the tapping positions for the control stimulus, a click signal of 5 ms duration, split by subject. Pooled over all subjects the taps precede the click signal by about 58 ms (SD 46.5 ms) which is in good agreement with findings in my former experiments and values found elsewhere. This effect of anticipation, known since Dunlop [22], was recently investigated in more detail by Radil et al. [23], Aschersleben/Prinz [24], and Gehrke [25].

Why anticipation occurs

One should not neglect the amount of estimations necessary to synchronise an action, the tapping of a hand, to a perceivable forthcoming acoustic event (in more detail Mates [30]). Figure 6 schematizes the timing relations of the involved contributors and gives an impression of what kind of estimates are necessary for a synchronized tapping task and why anticipation occurs.
The external world is only discoverable through our senses and the mental representations of sensible events. In figure 6 the external world is represented by the

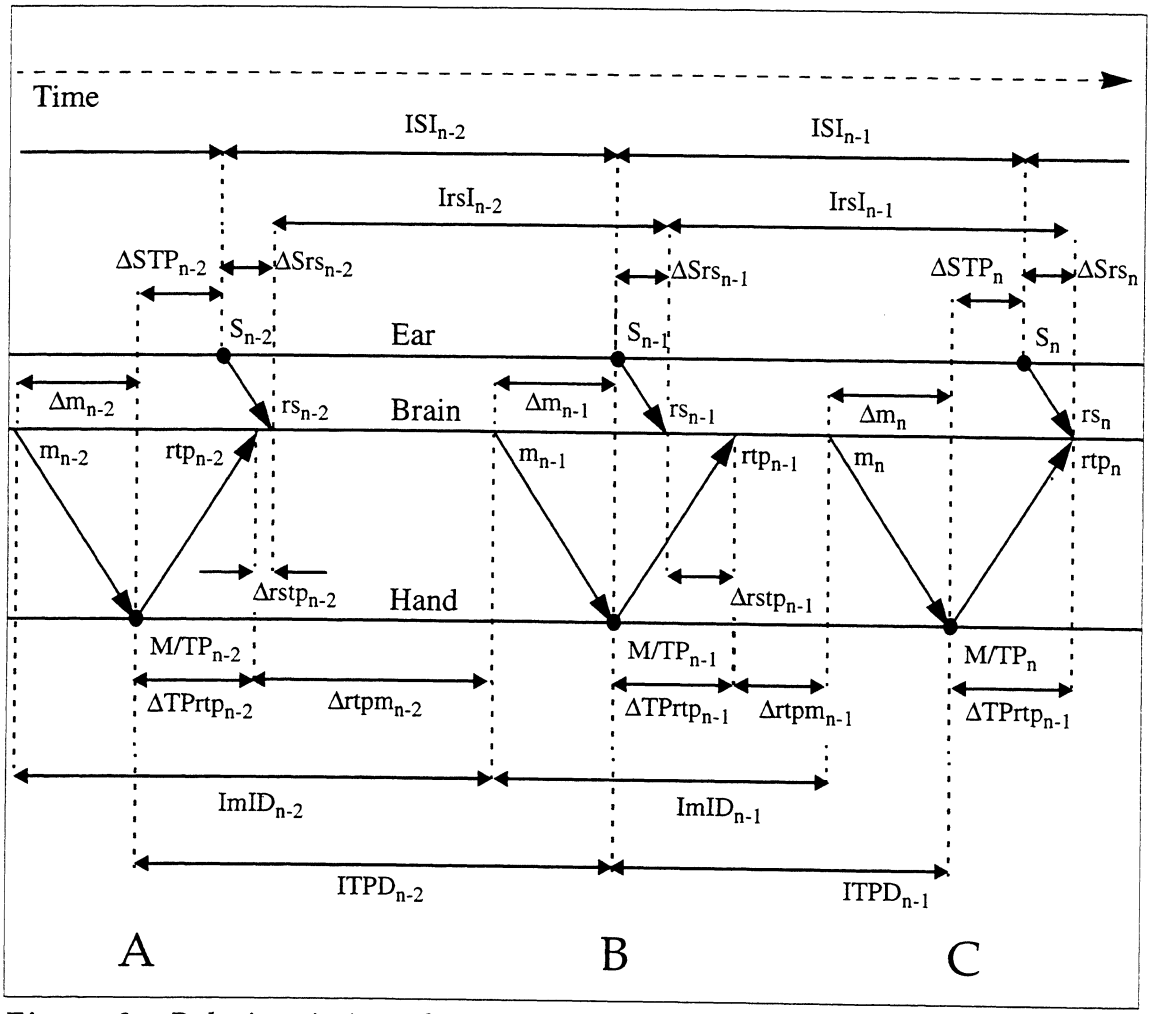

Figure 6: *Relative timing of external events (signal S; movement M; tap TP) and internal representations (rs, m, rtp) with estimated intervals (Δm, IrsI, ΔSrs, ΔTPrtp, Δrtpm, Δrstp, ImID), inter-tap distance(ITPD) and observable anticipation (ΔSTP).*

time event axes of ear and hand, the internal world by the time event axes of the brain (solid continuous lines). The ear-line is depicted closer to the brain, since the afferent information of the ear about a speech signal (S) reaches the brain earlier than the information of the hand about a tapping action (TP), as the ear is closer to the brain than the hand.

As we know subjects do not try to produce a similar time interval in between two events, but try to produce equally timed p-center alignments. Hence, they have to identify the inter-stimulus interval (ISI) of the signal sequence from the interval of the mental representations of successive perceived signals (IrsI). Furthermore to have a coincidence of a signal S and a tap TP (as in figure 6 case B) in the external world the subjects have to estimate the intervals ΔSrs (between the speech signal S and its representation rs), ΔTPrtp (between the tap TP and its representation rtp), Δrstp (the difference between these intervals) and Δm (between the mental representation of a movement instruction m and the actual movement M of the finger), to decide on the interval Δrtpm (between the last tap representation and the representation of the next movement instruction m) for the right moment in time to give the next movement instruction. A coincidence of signal S and tap TP

in the external world then occurs, if the subject is able to keep the interval Δrstp between the mental representation rs of the signal S and the feedback rtp of the tap TP, derived from the estimated intervals, constant at the magnitude of the subject specific difference ΔSrs minus ΔTPrtp. But the real world coincidence task does not only involve a lot of estimations, it also has to be done without any reliable feedback, as in the artificial situation of a tapping experiment an external validation of the coordination (like in tennis: hit or miss) of whether or not the task was performed successfully, is due to the experimental implications not present, and internal information for validation other than the estimates themselves is not available.
Real world events involving human action seem to coincide only if external information about the success of the performance is available and the performer able to compensate for the different time delays of the afferent and efferent information.
On the other hand, if we suppose that the task is accomplished by a coincidence of the mental representations rs and rtp of the signal and the tap (as in figure 6 case C) the subjects have only to make sure that there is no difference in time between the two representations. This can be done by adopting the IrsI interval as timing interval (ImID) for the movement instruction adjusted in a way that Δrstp gets minimized. This minimization process still involves a lot of estimation (Mates [30]) but no external information is necessary for validation; the task is successfully performed if the two event representations can no longer be differentiated in time. As a result the tap naturally has to precede the signal and a real world mismatch ΔSTP, the anticipation of the signal by the tapping subject, can be observed.

Normalization

Therefore, the anticipation measured above was also used to compensate for the effect of anticipation. That is, the tapping alignments of the subjects have been normalized by adding their individual anticipation[1], as measured in relation to the click signal, to the raw TP-values. This normalization reduces subject differences with respect to anticipation only, it does not alter any other individual differences like i.e. the amount of intra-subject variability, it especially does not change a possible stimulus effect. It also shifts the location of the overall mean tapping position by the amount of the mean anticipation into the positive direction with respect to the stimulus onset, that is further into the stimulus. Figure 7 gives the distribution of tapping responses for all subjects before and after normalization.
Out of the 29 subjects three seem to be unable to perform the task as intended. They show a much larger variation than the other subjects (figure 7 is somewhat misleading in that respect showing the responses pooled over all stimuli but the click). Figure 8 gives the 75% and 90% percentiles of the variances per stimulus for all subjects. There is a noticeable difference in the amount of variability for the three marked subjects in relation to all others. Their 75% percentiles are larger than the 80% percentile of the 90% percentiles of the other subjects. Furthermore

1) This does not assume that the 'individual' anticipation is an unalterable constant throughout the individuals entire life-span, but merely that it remains constant for the time of the experiment. If someone would participate in several experiments there would be an 'individual' anticipation value for every experiment.

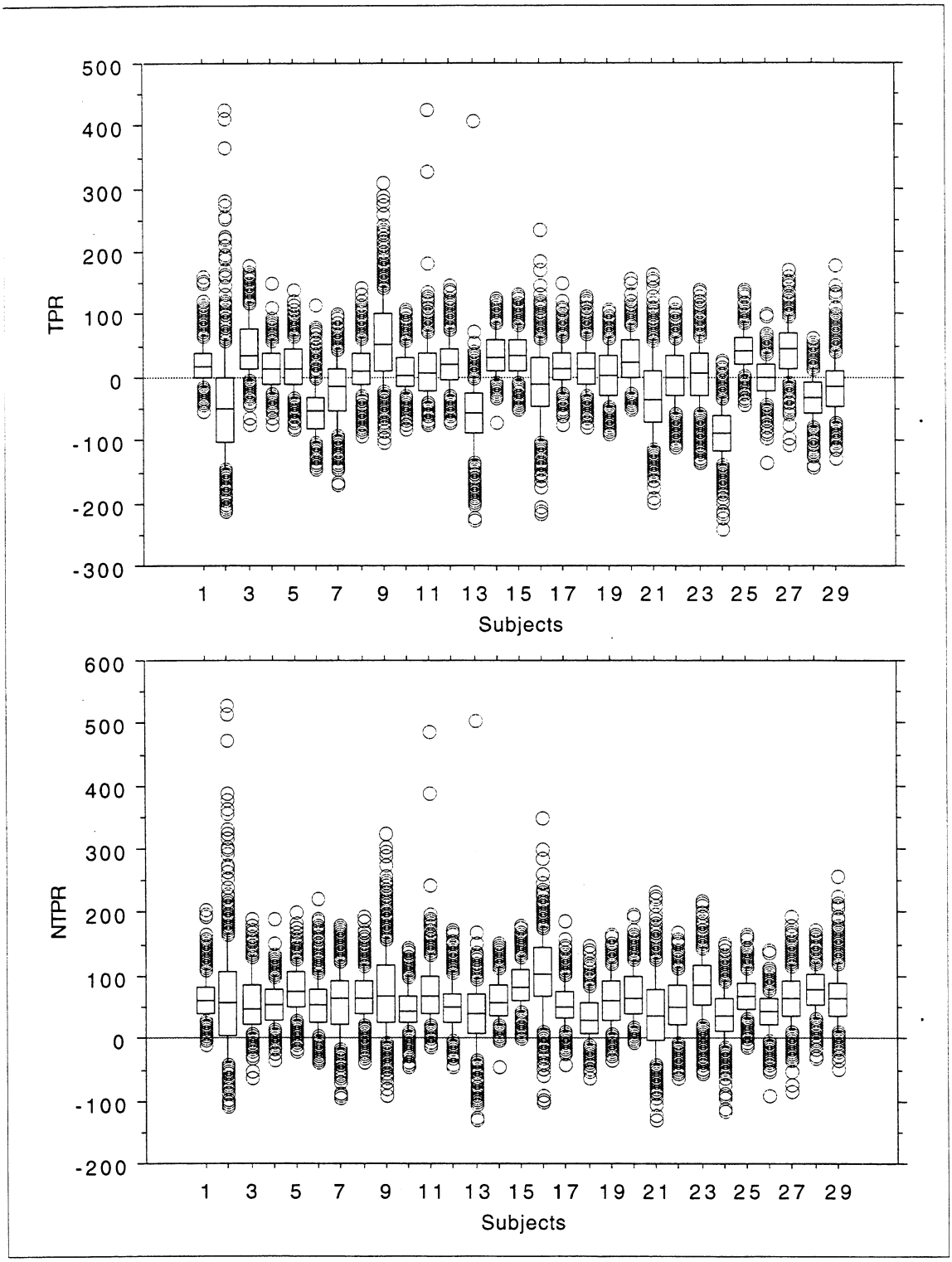

Figure 7: *Box plot of the tapping positions TPR and normalized tapping positions NTPR with median (middle line), inter quartile range (box), indicated 10% and 90% percentiles and outliers.*

they show significant differences (Scheffé (.01)) to other subjects in their inter quartile range, being not significantly different from one another, with the remaining 26 also not showing any significant difference in their inter quartile ranges. Figure 9 shows the 75% and 90% percentiles of the inter quartile range per

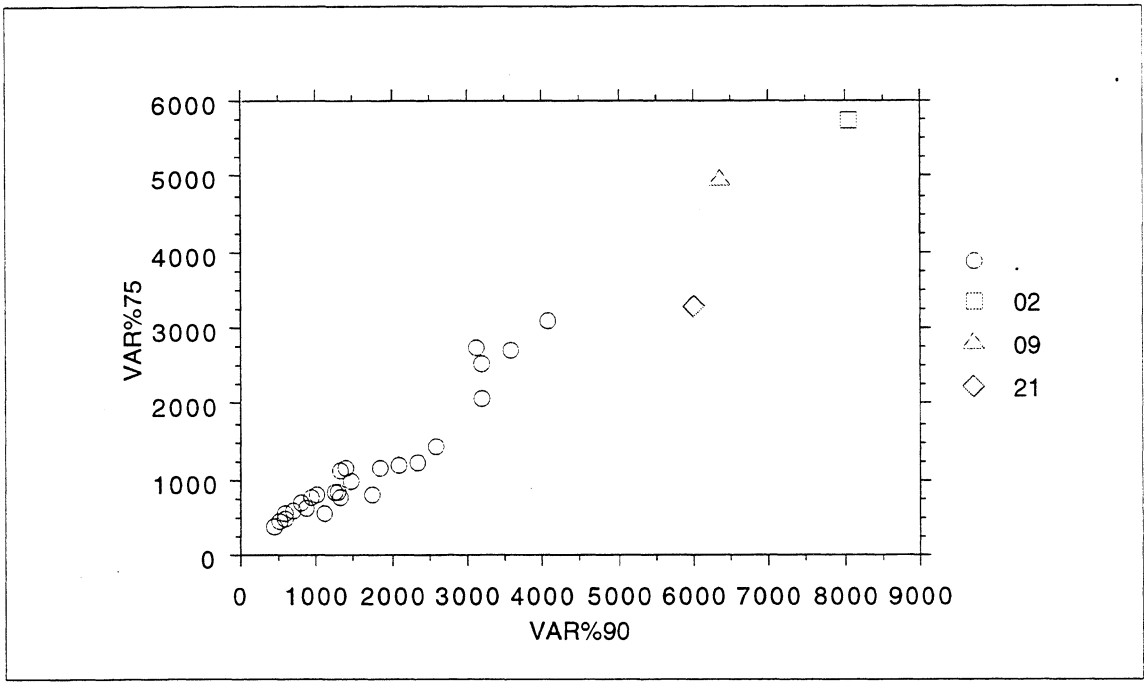

Figure 8: *75% and 90% percentiles of the variances per stimulus of the normalized tapping positions NTP for all subjects with three subjects marked showing a noticeable variability discrepancy.*

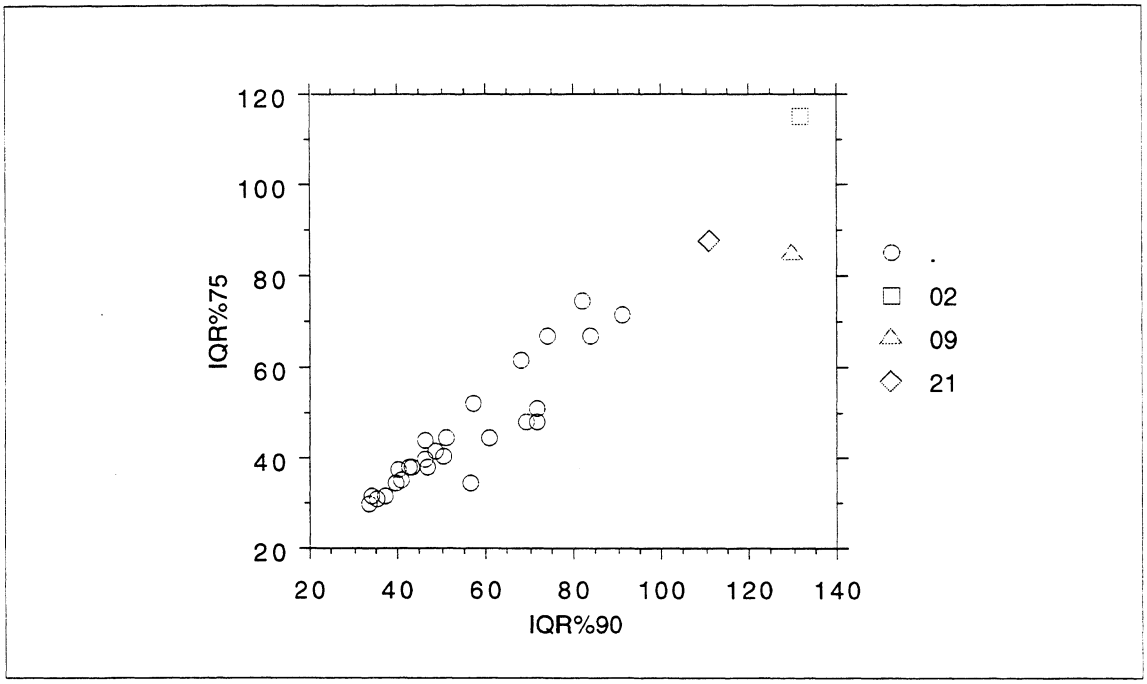

Figure 9: *75% and 90% percentiles of the inter quartile range per stimulus of the normalized tapping positions NTP for all subjects with three subjects marked showing a noticeable variability discrepancy.*

stimulus for all subjects, figure 10 the box plot of the inter quartile range. For further analysis the three marked subjects were omitted. Visual inspection seems to suggest the exclusion of one or two other subjects as well, but they did not come up in the analysis as being significantly different and remained therefore in the data set.

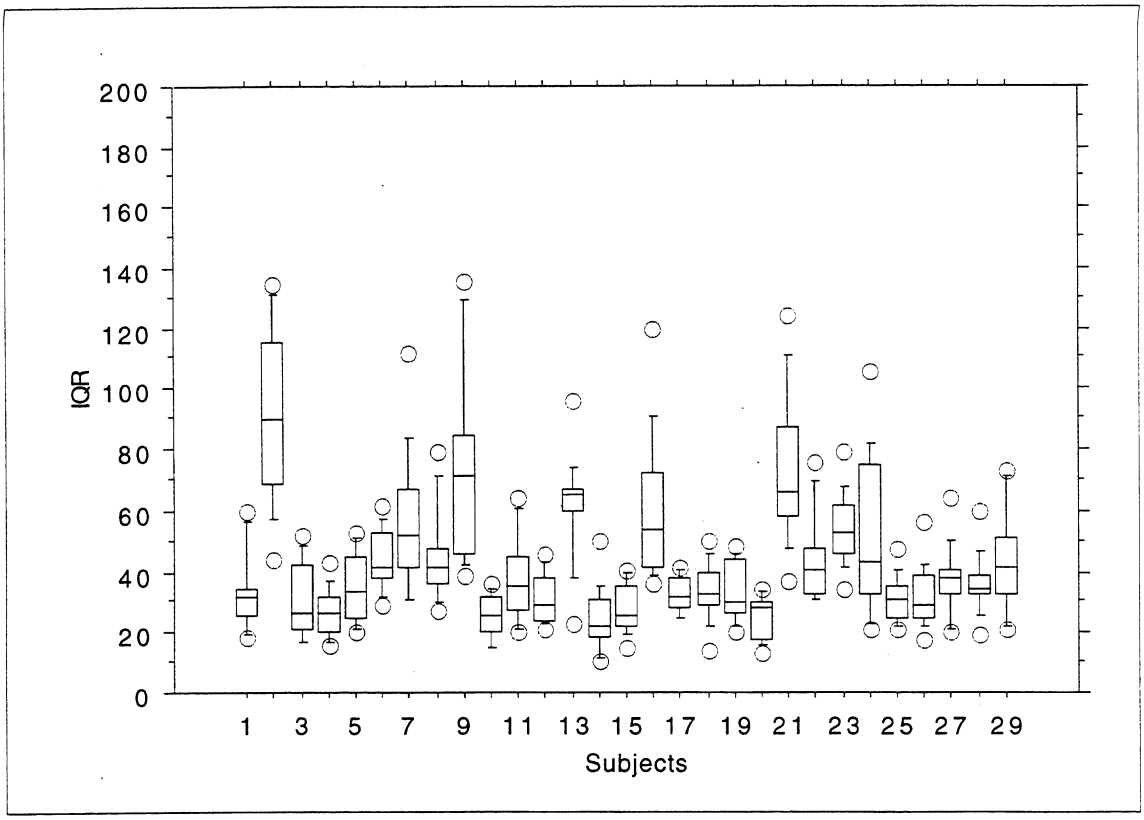

Figure 10: *Box plot of the inter quartile ranges IQR per stimulus for all subjects with median (middle line), inter quartile range (box), indicated 10% and 90% percentiles and outliers.*

Stimulus effects

For the remaining 26 subjects the repeated measurement analysis still gives, as expected due to the differences in the syllable initial consonants, a significant effect ($F = 114.58$, $p < .001$ / for words only $F = 115.41$, $p < .001$) for the factor stimulus as well as for the factor subject ($F = 23.03$, $p < .001$ / for words only $F = 18.89$, $p < .001$) with a significant interaction ($F = 1.46$, $p < .001$ / for words only not significant), due to the four already mentioned subjects which do not show a significant stimulus effect for the post hoc Scheffé (.01) test. Although the syllable rhyme has been kept phonotactically constant and the initial consonance has the main impact on the p-center, the significant differences for the factor stimulus are caused by the respective stimulus as a whole since the actual realization of the syllable rhyme differs physically. The locations of the normalized tapping positions NTP (with indicated standard deviation in the positive direction) can be seen in figure 11 in relation to the duration of the initial consonants, the vowel as syllable nucleus and the coda (see table I below). As can be seen the mean tappings are located around the consonant to vowel transition somewhat before and after the segmented syllable-nucleus-onset with no obvious standard deviation irregularities (mean sd =37.02 ms). Taken into account that the depicted locations are normalized for anticipation, this is in good agreement with the findings of my former experiments and the data reported in most of the above mentioned investigations.

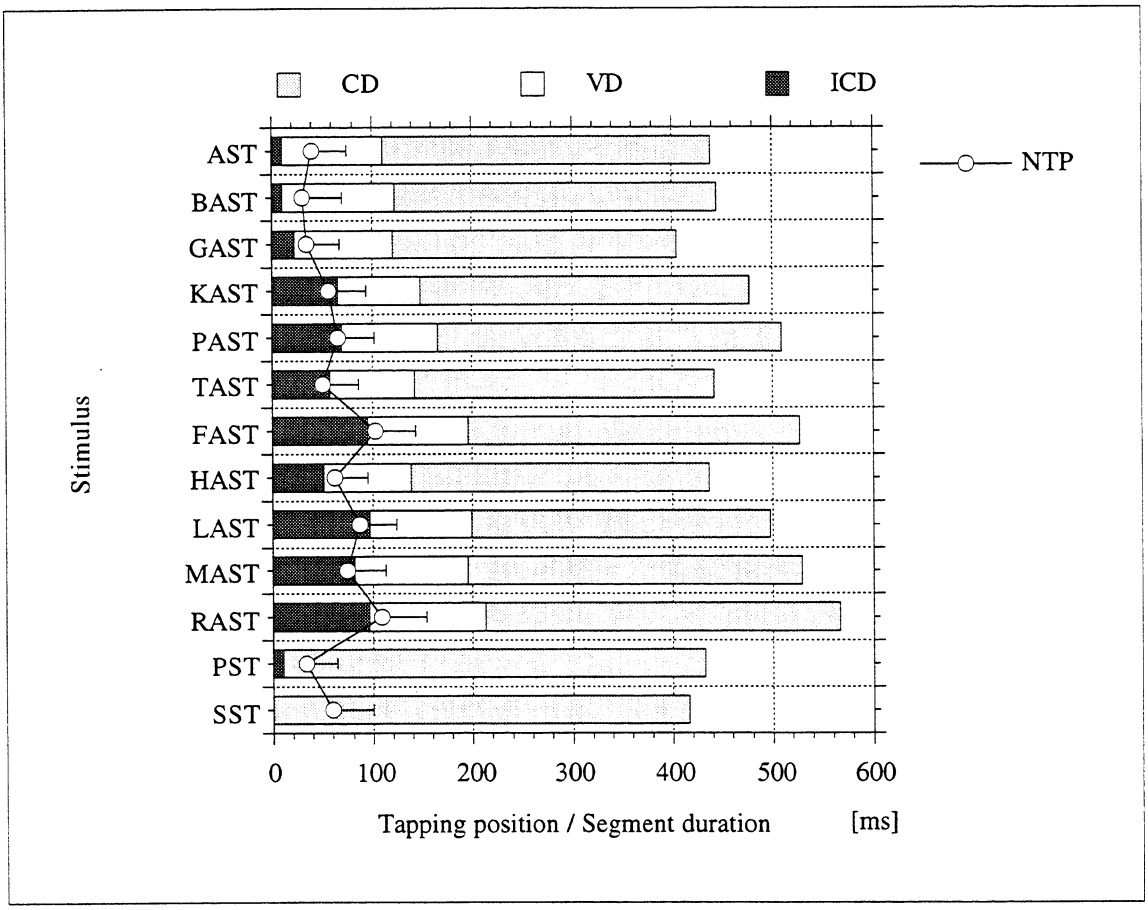

Figure 11: *Normalized tapping positions NTP for all stimuli, with indicated standard deviation in positive direction, in relation to the initial consonance duration (ICD = syllable-nucleus-onset), nucleus duration (VD) and coda duration (CD).*

Model estimates and correlations

To assess, how good the measured syllable-nucleus-onset as a predictor of the p-center alignment fits with the location of the normalized tapping positions and to compare it to other estimates of the p-center position for some of the above mentioned models their estimates and the corresponding correlations were computed. Table I gives the measured values as well as the model specific estimates for the models of Schütte (SHT), Marcus (MPC), Köhlmann (KLM), Howell (HWL) and Pompino-Marschall (BPM), table II the corresponding correlation matrix. The correlations between the normalized tapping positions and the model estimates as well as between the normalized tapping positions and the syllable-nucleus-onset are quite high (Fisher's r to z cor > .827, $p < .0009$). As expected the syllable nucleus, coda or rhyme do not correlate with either the tappings or any of the models, but the rhyme (nucleus + coda) correlates more highly with the coda ($p < .0001$) than with the nucleus ($p = .0084$) showing that the nucleus has to be more stable while the coda is more variable that is more open for compensational effects. Figures 12 and 13 show the correlations between the normalized tapping positions and the model estimates for the word stimuli as well as the syllable-nucleus-onset. The correlations are only calculated for the word stimuli as some of the models rely on entities not available with the interjections.

Table I: Measurements and model estimates

Normalized tapping positions NTP with standard deviation SD, measured duration of initial consonance ICD, nucleus VD, coda CD, syllable rhyme RM and estimated alignment values for the models of Marcus MPC, Pompino-Marschall BPM, Howell HWL, Schütte SHT, and Köhlmann KLM.

Stimulus	NTP	SD	ICD	VD	CD	RM	MPC[1]	BPM	HWL	SHT	KLM
AST	39.54	35.3	10	101	328	429	113.8	14.5	68.8	12.0	10.3
BAST	30.16	39.7	10	113	322	435	115.2	15.1	64.4	12.9	9.1
GAST	34.40	33.3	23	98	284	382	110.5	28.6	82.6	27.4	13.7
KAST	56.11	36.9	66	83	328	411	145.7	61.1	117.	60.0	34.0
PAST	65.28	36.5	70	96	343	439	155.2	64.2	117.5	73.3	35.3
TAST	50.44	35.3	58	85	299	384	133.7	51.0	109.3	60.4	33.6
FAST	102.70	40.2	102	94	331	425	172.5	104.6	152.6	106	47.0
HAST	62.36	32.2	59	80	298	378	132.9	48.9	104.8	60.2	26.1
LAST	87.03	36.8	90	109	298	407	160.3	77.9	147.2	46.4	58.1
MAST	74.12	38.2	82	113	334	447	165.1	74.6	137.4	40.8	51.7
RAST	108.30	45.2	96	117	354	471	180.2	110.2	174.3	98.3	58.2
PST	32.81	31.5	6	0	422	422	109.4	33.9	122.6	15.8	8.4
SST	59.78	40.1	0	0	416	416	104.0	69.2	123.4	41.3	39.6

[1] Can be distinctly improved in relation to alignment location depending on a post hoc optimization of the so-called 'arbitrary' constant (see text)

Table II: Correlation matrix for measurements and estimates

Normalized tapping positions NTP for words, duration of initial consonance ICD, nucleus VD, coda CD, syllable rhyme RM, estimates of the models of Marcus MPC, Pompino-Marschall BPM, Howell HWL, Schütte SHT, Köhlmann KLM, p-values for the framed correlations: SHT/NTP $p = .0008$, SHT/ICD $p = .0005$, all others $p < .0001$

	NTP	ICD	MPC	BPM	HWL	SHT	KLM	VD	CD	RH
NTP	1.000	.932	.949	.969	.963	.828	.907	. .243	.458	.446
ICD	.932	1.000	.957	.969	.969	.842	.952	.048	.345	.277
MPC	.949	.957	1.000	.970	.966	.805	.946	.262	.577	.543
BPM	.969	.969	.970	1.000	.986	.883	.927	.164	.456	.410
HWL	.963	.969	.966	.986	1.000	.819	.969	.215	.408	.397
SHT	.828	.842	.805	.883	.819	1.000	.688	-.180	.415	.230
KLM	.907	.952	.946	.927	.969	.688	1.000	.271	.347	.376
VD	.243	.048	.262	.164	.215	-.180	.271	1.000	.397	.731
CD	.458	.345	.577	.456	.408	.415	.347	.397	1.000	.916
RH	.446	.277	.543	.410	.397	.230	.376	.731	.916	1.000

The simple threshold model – compared to some of the other models – of Schütte, taking overall duration and amplitude envelope into account, shows at .828

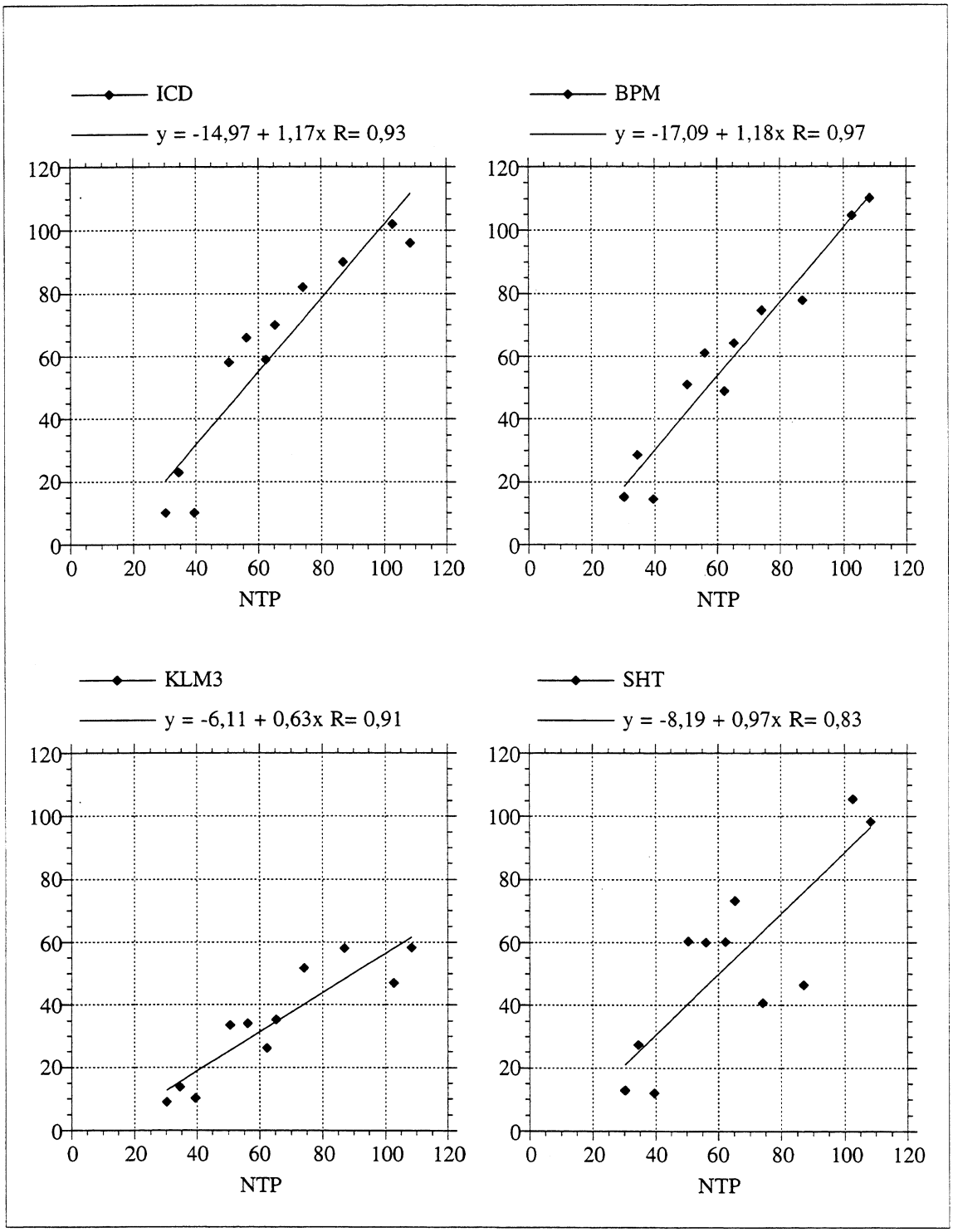

Figure 12: *Correlation and regression line of the normalized tapping positions NTP and the measured syllable-nucleus-onset ICD as well as the model-derived estimations of the alignment position of the word stimuli for the models of Pompino-Marschall (BPM), Köhlmann (KLM3) and Schütte (SHT).*

(p = .0008) the least but still a good correlation to the tappings. The correlation between the normalized tapping positions and the syllable-nucleus-onset is at .932 quite high, but not as good as the correlation of the original p-center model by

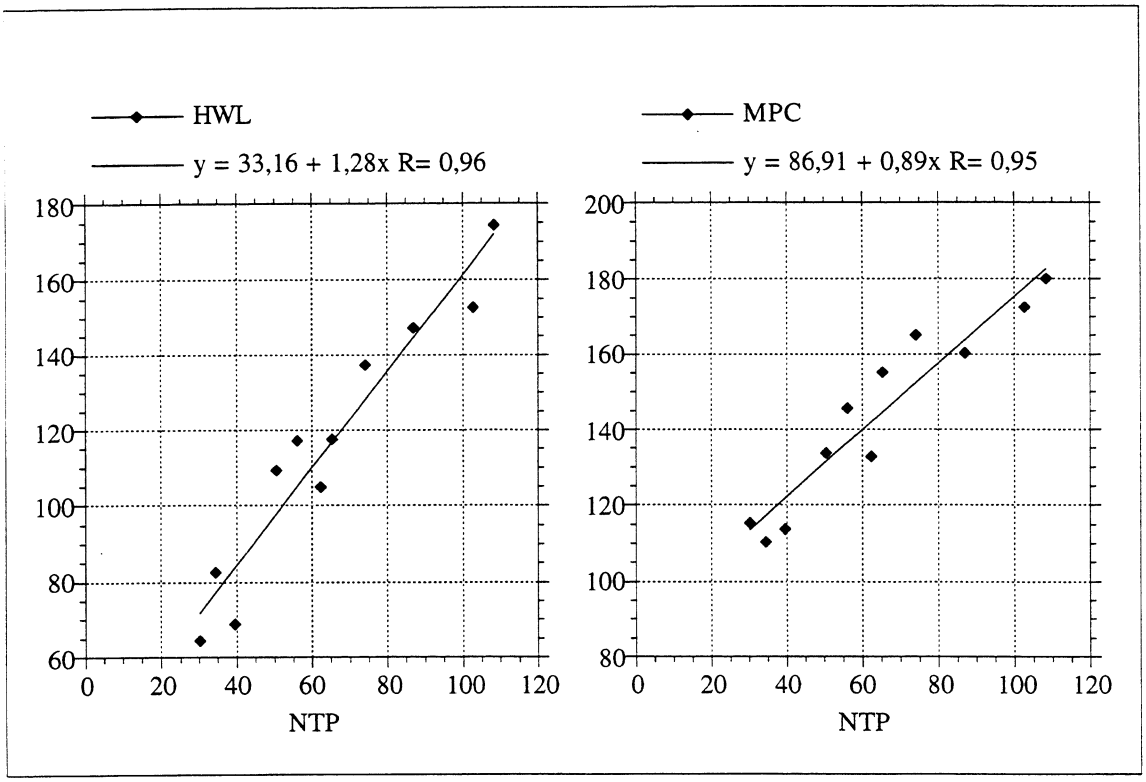

Figure 13: *Correlation and regression line of the normalized tapping positions NTP and the model-derived estimations of the alignment position of the word stimuli for the models of Howell (HWL) and Marcus (MPC) (note the difference in the y-axes values).*

Marcus, which weights the duration of the initial consonance and the rhyme. This is somewhat surprising since a correlation of the rhyme with the tapping positions (Fisher's r to z cor = .446, n.s.) could not be found.
The psycho-acoustic model of Köhlmann uses the pitch and the loudness contours of the acoustic signal to detect onset events, which are integrated to form an 'Ereigniszeitpunkt' (EZP, moment of occurrence) if closer in time to one another than 120 ms. This sometimes leads to more than one event (EZP) per stimulus. For the correlation analysis here – as this is done to assess the predictive ability of the p-center syllable-nucleus-onset hypothesis, not to discuss the problems of the Köhlmann model – if in doubt the events which give the better correlation are used. At .907 the correlation is not as good as the correlation of the nucleus onset.
The psycho-acoustic model of Pompino-Marschall gives at .969 the best correlation. This model calculates partial onset and offset events for rising and falling flanks of the loudness contour of the single critical bands and integrates them to a single 'syllable onset'.
At .963 the p-center estimates of Howells model – which can be interpreted for monosyllabic signals as the area bisect of the rectified signal amplitude envelope – also correlate extremely well.
Unfortunately, a high correlation as such is not sufficient to assess the predictive strength of a model and the reliability of the model estimates.
As can be seen in figure 14, which depicts the model estimates of the four best correlating models (table II) in relation to the categorical segments of the stimuli, most of the estimates do monitor the tendency of the differences of the normal-

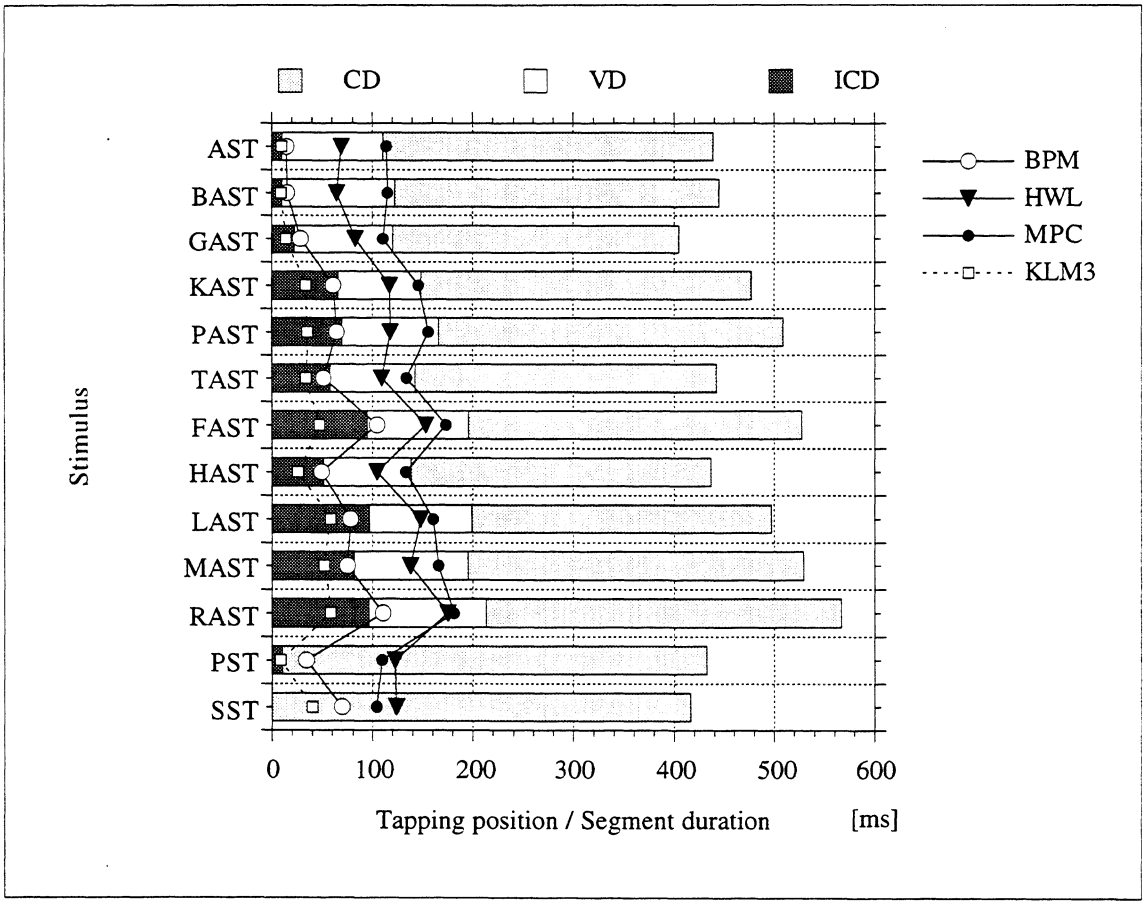

Figure 14: *The estimated alignment positions of all stimuli for the models of Pompino-Marschall (BPM), Howell (HWL), Marcus (MPC) and Köhlmann (KLM3) in relation to the segments of the stimuli (duration of ICD = initial consonance (syllable-nucleus-onset), VD = nucleus, CD = coda).*

ized tapping position quite well but only the BPM estimates are for all stimuli in the magnitude of the normalized tapping positions and hence close to their actual location. For some of the stimuli also the SHT and the KLM models give estimates in the magnitude of the tappings. The magnitude of the MPC estimates can be distinctly improved with respect to the place of the alignment, if one chooses an appropriate constant factor. The problem here simply is, that one needs to know the alignment position beforehand to choose the 'optimal' one (for this data cons. = 76.56). Marcus does not provide any, as the model is not intended to give absolute[1] p-center estimates, and the one I have found for other stimuli elsewhere (cons. = 17.02) wouldn't have helped with the data reported here.
Furthermore, as one might have already noticed and can be seen in figure 15, the correlation matrix (table II) also reveals that all of the models correlate at least as good as or even more highly with the syllable-nucleus-onset than with the normalized tapping positions. In this respect it seems, that the different models used to estimate the alignment position are just some very sophisticated methods to estimate the variation and trend of the syllable-nucleus-onset.

---

1) Marcus [11] states that the constant "is an arbitrary constant representing the fact that we are only determining *relative* (emphasis by Marcus) P-center locations of stimuli to one another." (p. 252f).

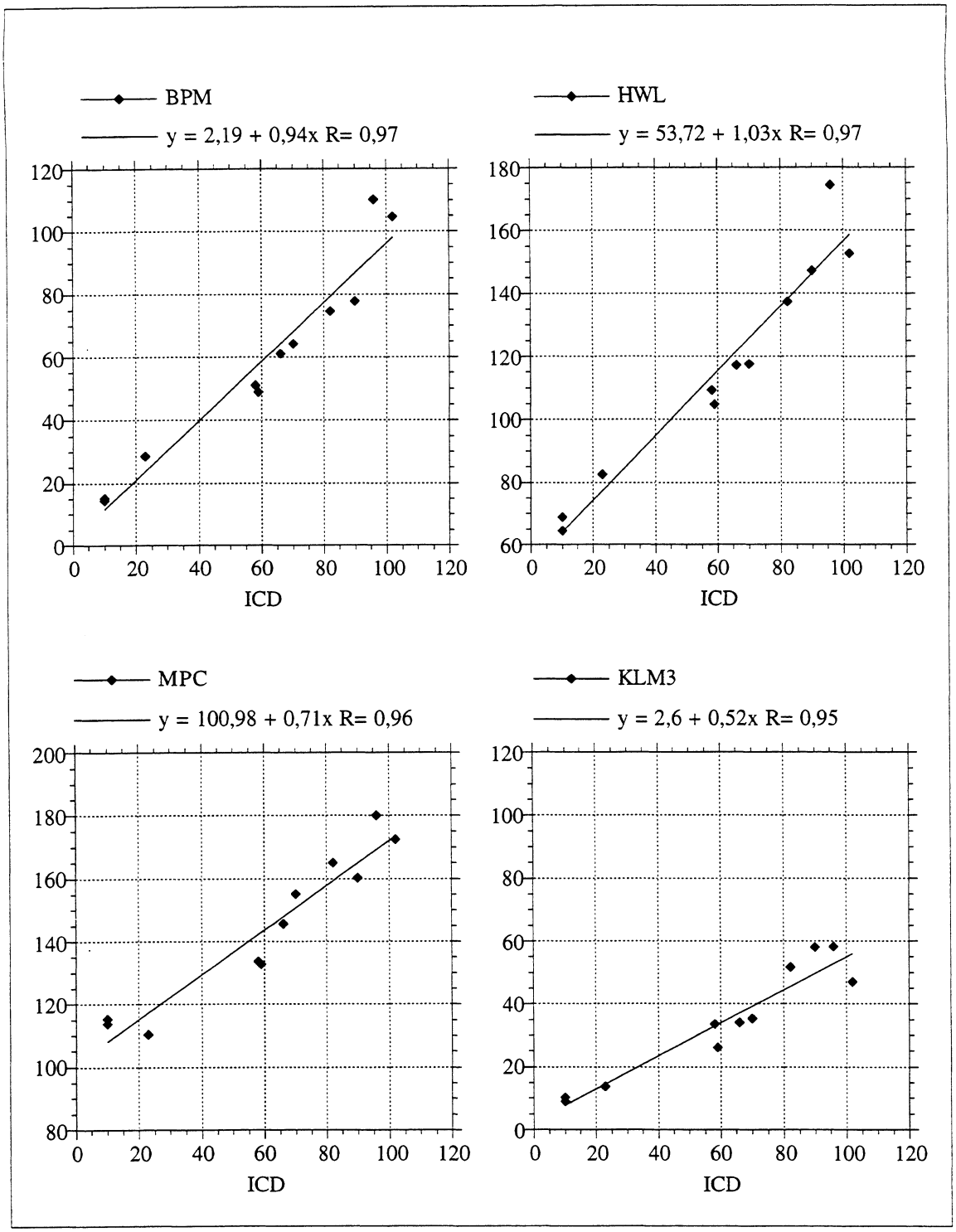

Figure 15: *Correlation of the measured syllable-nucleus-onset ICD and the model-derived estimations of the alignment position of the word stimuli for the models of Pompino-Marschall (BPM), Howell (HWL) Marcus (MPC) and Köhlmann (KLM3).*

This result suggests that there is good evidence for the hypothesis of a p-center syllable-nucleus-onset correspondence but a complete match could not be found. The question therefore is, what might be the cause for this mismatch here and the mismatches found elsewhere.

• Firstly the anticipation phenomenon as described above.

The fact, that in most of the other investigations the location of the alignment was somewhat before the consonance-vowel transition was already mentioned above, and might be at least partly due to this phenomenon for which the data here has been corrected. As the method used here is quite straight forward but imprecise, there might be still some inaccuracy due to anticipation, but in my opinion not a substantial one.

• Secondly, and more important the measurements for the segment boundaries have a large influence on the match.

Some variability and uncertainty is introduced by the fact that, depending on the rules for segmentation in use, and the clarity of the signal, a difference of one or two glottal pulses in the location of the segment boarder in both directions can easily occur. Depending on the fundamental frequency the speaker is speaking with, that is, the period duration (70 Hz ≈ 14 ms/100 Hz ≈ 10 ms) of the signal, this can account for up to 50 ms difference for the measured onset in the data. Furthermore, whether this measured postulated signal syllable-nucleus-onset really represents the mental representation of the syllable-nucleus-onset with sufficient accuracy is also not completely clear. Unfortunately this kind of variability and uncertainty cannot be averted.

• Thirdly, the ability to perform the intended task successfully.

As already mentioned above in experiments with simultaneous tapping to a presented signal, subjects have to perform several tasks at once: recognizing the rhythm pattern, tapping this pattern by estimating the occurrence of the next events in time, judging whether or not they matched the presented event with the tapping action. Hence, quite large difference in subject responses can be expected. The consideration of misleading measurements caused by the first two aspects can partly be avoided as the inability of recognizing the rhythm pattern, estimating the next event or coordinating the tapping action manifests itself for the subjects concerned in a noticeable larger variability of the responses not between stimuli but within the respective stimulus.

The third aspect, judging the co-occurrence, is limited by the fusion and the order threshold, the ability to recognize differences between two successive events (such as Δrstp in figure 6 case A) and the order in which they occur. For auditory signals as such this order threshold (Pöppel [27], Steinbüchel/Pöppel [28]) is supposed to be at about 20 ms to 35 ms (fusion threshold 2 ms), depending on the measurement procedure (recently discussed in detail by Steffen [29]).

Allen [15] found in his data, where a click signal was superimposed on a speech signal, that a range of about 200 ms seems to be an interval within which his subjects still judged events as being simultaneous. Compared to the estimates of the models and hypotheses discussed here – all of them would suggest alignment relations within this distance from the location of the tapping positions and the consonance-vowel transitions – these findings are rather vague and can not be a basis to decide whether any of the models and hypotheses is more successful than the other in estimating the alignment relation.

Therefore, the order threshold of about 20 ms was used as a basis to define an uncertainty or indifference interval (II) of 40 ms within which the order of two events, i.e. tapping and presented syllable, cannot doubtlessly be told. The number of taps within this interval around the estimates of the various models was counted and fed into a contingency analysis (see table III). The observed fre-

Table III: Model specific indifference interval contributions

Contingency Analysis with observed frequencies, expected values and percents of column totals for the indifference intervals of the syllable-nucleus onset ICDII and the estimates of the models of Marcus[1] MPCII, Pompino-Marschall BPMII, Howell HWLII, Schütte SHTII, Köhlmann KLMII (Chi Square = 7315.171, p < .0001)

		ICDII	BPMII	KLMII	MPCII[1]	HWLII	SHTII	Totals
observed frequencies	outside	7425	7669	8931	12161	10798	7719	54703
	inside[2]	5335	5091	3829	599	1962	5041	21857
	totals	12760	12760	12760	12760	12760	12760	76560
Expected Values	outside	9117.167	9117.167	9117.167	9117.167	9117.167	9117.167	54703
	inside	3642.833	3642.833	3642.833	3642.833	3642.833	3642.833	21857
	totals	12760	12760	12760	12760	12760	12760	76560
Column Totals (%)	outside	58.190	60.102	69.992	95.306	84.624	60.494	71.451
	inside[3]	41.810	39.898	30.008	4.694[1]	15.376	39.506	28.549
	totals	100	100	100	100	100	100	100

1) Can be distinctly improved by post hoc optimization of the so-called 'arbitrary' constant (see text above).

2) The maximum value for this row is the number of tappings inside the indifference interval NTPII of the normalized tapping positions NTP which is 5666.

3) The maximum value for this row is the %-inside value for the indifference interval NTPII of the actual data which is 44.4 %.

quency distribution confirms the mentioned differences in magnitude of the estimations, as the two models (HWL, MPC) with estimations further away from the tapping locations show low to very low values for tappings inside the indifference intervals around their estimates with only about 15.4 % of the tappings inside the indifference interval HWLII for the HWL model and hardly any, that is about 4.7 % inside MPCII for the original p-center model of Marcus. For KLMII about 30 % are reported, which is slightly more than the expected value of 28.5 % under the assumption that there is no difference between the models and the overall distribution between values inside and outside the indifference interval is as given by the entire models. Assuming the indifference interval at the mean of a normal distribution with the standard deviation of the actual tapping data, the expected percentage of values within the indifference interval would be 40 %. This is about the value the analysis offers for the models of Schütte and Pompino-Marschall. Remembering the findings for the correlation analysis, of the computational models the psycho-acoustic model of Pompino-Marschall clearly is the model of choice to determine an alignment position, giving an estimate in mag-

nitude close to the actual tapping locations for words as well as for interjections and showing a very good correlation with the variation of the alignment position introduced by the stimuli.
The only estimate that gives a noticeable higher amount of tappings inside the indifference interval is the syllable-nucleus-onset at about 41.8 %. That is just about 2% below the experiment specific maximal value of 44.4 % and nearly 2% more than could have been expected from a normal distribution with this standard deviation, which indicates that the data is slightly squeezed into the middle of the distribution (positive kurtosis) and that the syllable-nucleus-onset is the alignment marker which accounts best for that fact.
Thus for the words presented here, the estimations derived from the measured syllable-nucleus-onset give the overall best prediction of the alignment position.

CONCLUSION

The results clearly show that the measured syllable-nucleus-onset is at least an equally good estimate for the location of the alignment position (here tapping) for equally timed rhythm perception as any of the other mentioned models of estimation, and therefore supports strongly the notion of a p-center syllable-nucleus-onset correspondence hypothesis or better, a congruence of the p-center and the mental representation of the syllable-nucleus-onset.

ACKNOWLEDGEMENTS

This work was supported by a grant of the "Forschungsschwerpunkt Allgemeine Sprachwissenschaft, Typologie und Universalienforschung der Förderungsgesellschaft Wissenschaftliche Neuvorhaben mbH Berlin".

REFERENCES

[1] TERHARDT, E.; SCHÜTTE, H. (1976), Akustische Rhythmuswahrnehmung: Subjektive Gleichmäßigkeit. *Acustica* 35, pp. 122-126.

[2] MORTON, J., MARCUS, S.M., & FRANKISH. C. R. (1976), Perceptual centres (P-centers), *Psychological Review,* vol. 83, pp. 405-408.

[3] MARCUS, S. M. (1976), *Perceptual Centres.* unpublished PhD thesis, Cambridge University Cambridge, England.

[4] HOWELL, P. (1988), Prediction of p-center location from the distribution of energy in the amplitude envelope: Part I & II, *Perception & Psychophysics,* vol. 43, pp. 90-93 & 99.

[5] JANKER, P. M. (1989), Der Einfluß von Segmentdauer- und Amplitudenmanipulation auf die P-center-Position einfacher CV-Silben, *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München,* vol. 27, pp. 71-141.

[6]JANKER, P. M. (1995), Sprechrhythmus, Silbe, Ereignis. Eine experimentalphonetische Untersuchung zu den psychoakustisch relevanten Parametern zur rhythmischen Gliederung sprechsprachlicher Äußerungen, *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, vol. 33, pp. 1-260.

[7]JANKER, P. M. (1995), On the influence of the internal structure of a syllable on the p-center perception, *Proc. 13th ICPhS Stockholm*, vol. 2, pp. 510-513.

[8]JANKER, P. M., POMPINO-MARSCHALL, B. (1991), Is the p-center influenced by 'tone'?, *Proc. 12th ICPhS Aix-en-Provence*, vol. 3, pp. 290-293.

[9]JANKER, P. M., The range of subjective simultaneousness in tapping experiments with speech stimuli, *Proc. of the ESCA Workshop on the Auditory Basis of Speech Perception*, Keele (UK) 15.-19. Juli 1996, pp. 204-207.

[10]KÖHLMANN, M. (1984), Rhythmische Segmentierung von Sprach- und Musiksignalen und ihre Nachbildung mit einem Funktionsschema, *Acoustica*, vol. 56, pp. 192-204.

[11]MARCUS, S. M.(1981), Acoustic determinants of perceptual center (P-center) location, *Perception & Psychophysics*, vol. 30(3), pp. 247-256.

[12]POMPINO-MARSCHALL, B. (1989), On the psychoacoustic nature of the P-center phenomenon, *Journal of Phonetics*, vol. 17, pp. 175-192.

[13]POMPINO-MARSCHALL, B. (1990), *Die Silbenprosodie. Ein elementarer Aspekt der Wahrnehmung von Sprechrhythmus und Sprechtempo*, Tübingen: Niemeyer.

[14]SCHÜTTE, H. (1978), Ein Funktionsschema für die Wahrnehmung eines gleichmäßigen Rhythmus in Schallimpulsfolgen, *Biological Cybernetics* 29, pp. 49-55.

[15]ALLEN, G. D. (1972), The location of rhythmic stress beats in English: An experimental study I & II. *Language and Speech* 15, pp. 72-100 & 179-195.

[16]RAPP, K. (1971) A study of syllabic timing, *Speech Transmission Lab., Royal Inst. Technology Quarterly Status and Progress Report 1*, pp. 14-19 Stockholm.

[17]HOLLISTER, R. D. T. (1973), Relation between hand and voice impulse movements, *Speech Monogr.* 4 (1), pp. 75-100.

[18]MEYER, E. A. (1898), Beiträge zur deutschen Metrik. Über den Takt, *Die neueren Sprachen*, vol. 6, pp. 1-37 & 122-140.

[19]JANKER, P. M. (1994), Nature and properties influencing the p-center perception. RPP94, Sheffield 8.-11. September 1994.

[20]JANKER, P. M. (1996), More on the nature and properties influencing the p-center perception. RPP96, Ohlstadt 8.-11. September 1996.

[21]JANKER, P. M. (1996, to appear), Some evidence for the p-center syllable-nucleus-onset correspondence hypothesis.

[22]DUNLAP, K. (1910), Reactions on rhythmic stimuli, with attempt to synchronize, *Psychological Review*, vol. 17, pp. 399-416.

[23]RADIL, T., MATES, J., ILLMBERGER, J., PÖPPEL, E. (1990), Stimulus anticipation in following rhythmic acoustical patterns by tapping, *Experimentia*, vol. 46, pp. 762-763.

[24]ASCHERSLEBEN, G., PRINZ, W. (1992), What gets synchronized with what in sensorimotor synchronisation, *Paper 13/1992, MPI für psychologische Forschung*, München.

[25]GEHRKE, J. (1996), Presentation at the RPP96, Ohlstadt 8.-11. Sep. 1996.

[26]ASCHERSLEBEN, G. (1994), *Afferente Information und die Synchronisation von Ereignissen*, Frankfurt/M: Lang.

[27]PÖPPEL, E. (1978), Time perception. Held, R., Leibowitz, H., Teuber, H.-L.: *Handbook of sensory physiology, vol. VIII: Perception*, pp. 713-729, Berlin: Springer.

[28]STEINBÜCHEL, N.v., PÖPPEL, E. (1991), Temporal order threshold and language perception. Bhathkar, V., Rege, K.: *Frontiers in knowledge-based computing*. New Delhi: Narosa.

[29]STEFFEN, A. (1995), *Zur Bestimmung der Ordnungsschwelle: Ein experimenteller Vergleich,* Magisterarbeit LMU München.

[30]MATES, J. (1992) Timing and corrective mechanisms in synchronization of motor acts with a sequence of stimuli, *Proc. Fourth Rhythm Workshop: Rhythm Perception and Production,* pp. 43-48 Bourges.

SUPPLEMENT

The 10 stimuli of the experiment not presented in the stimulus section.

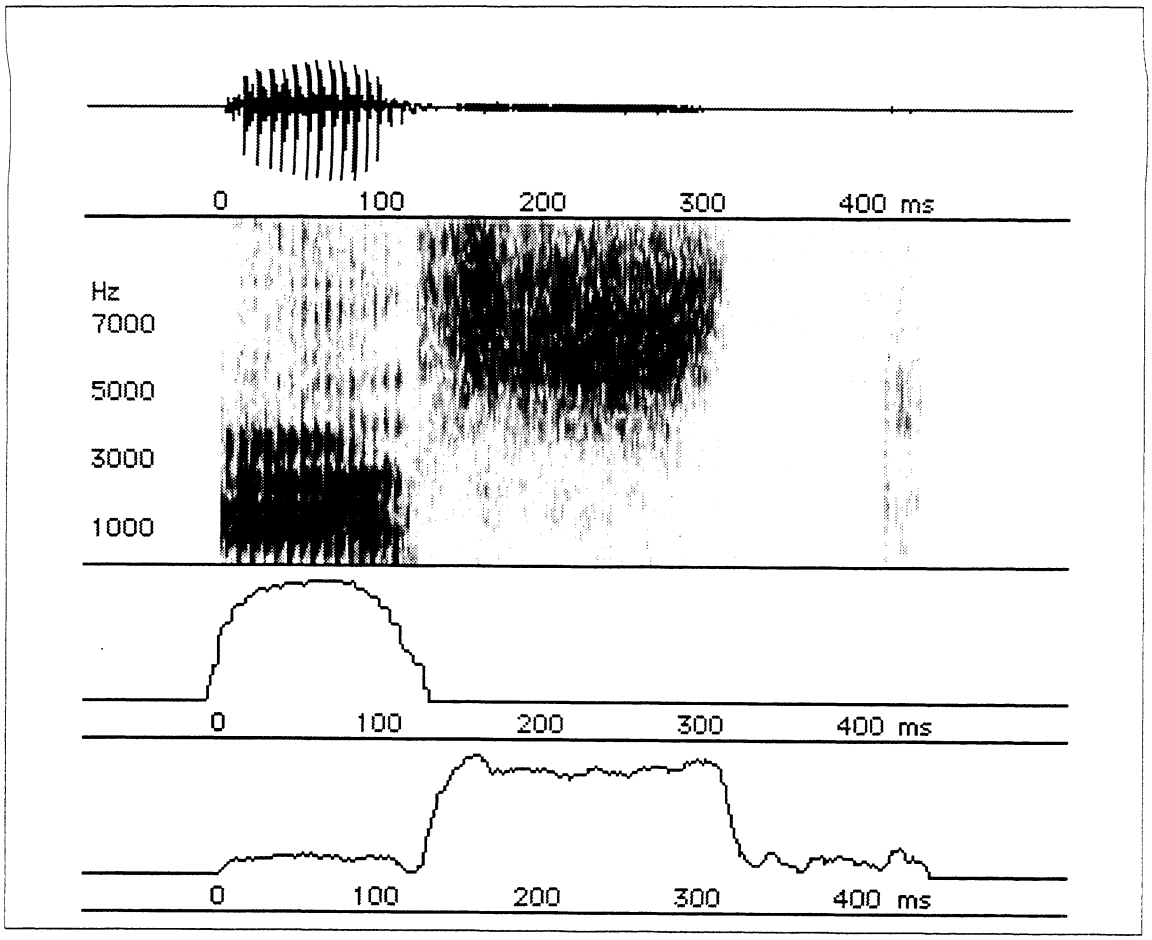

Figure 16: *<ast> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <ast> with an overall duration of 439 ms.*

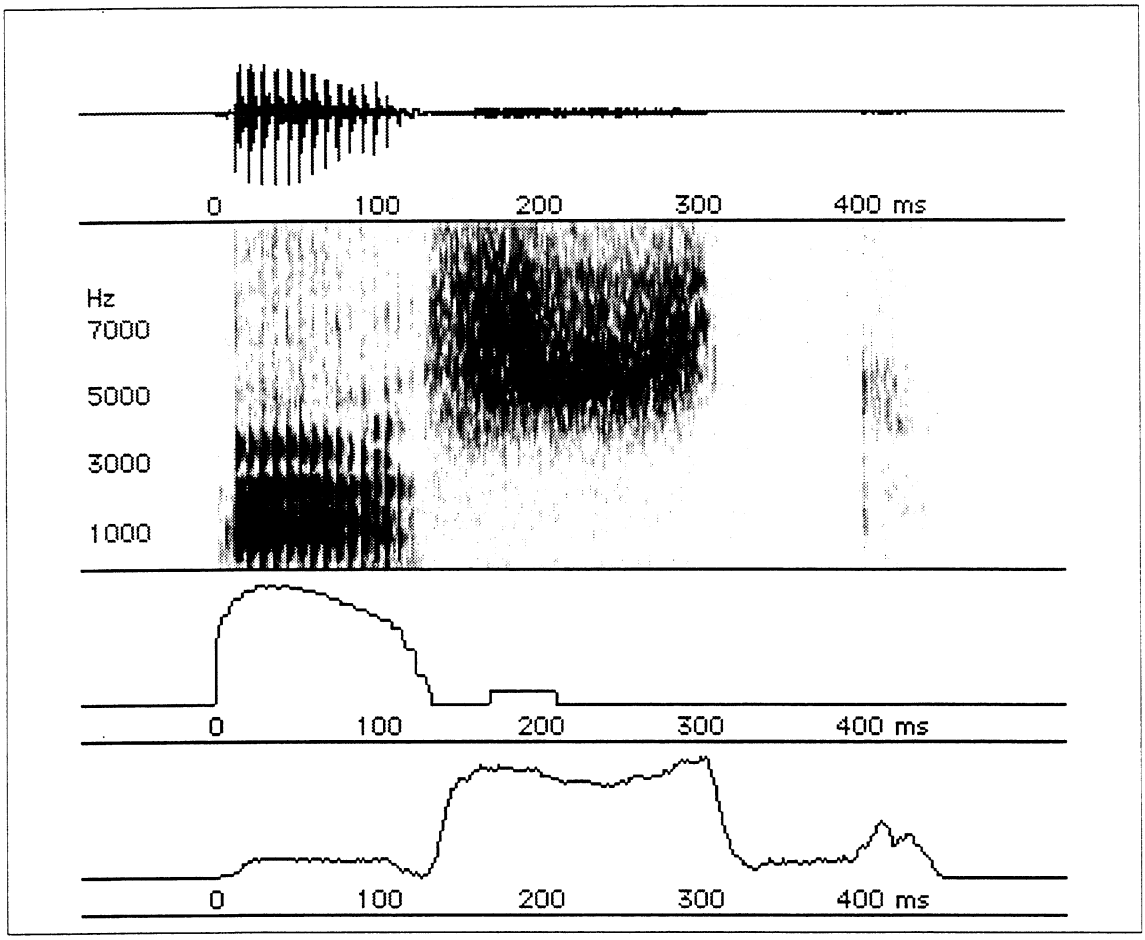

Figure 17: *<bast> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <bast> with an overall duration of 445 ms.*

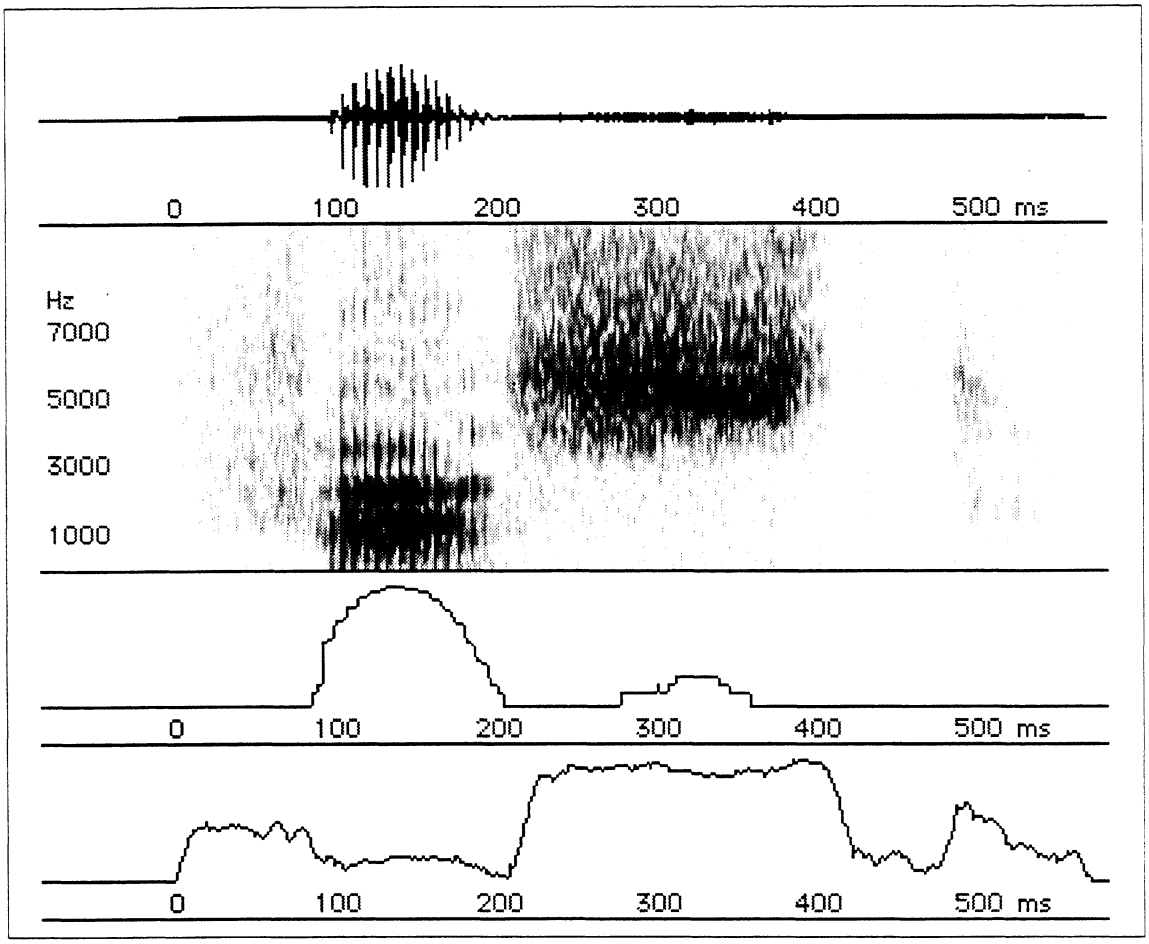

Figure 18: *<fast> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <fast> with an overall duration of 527 ms.*

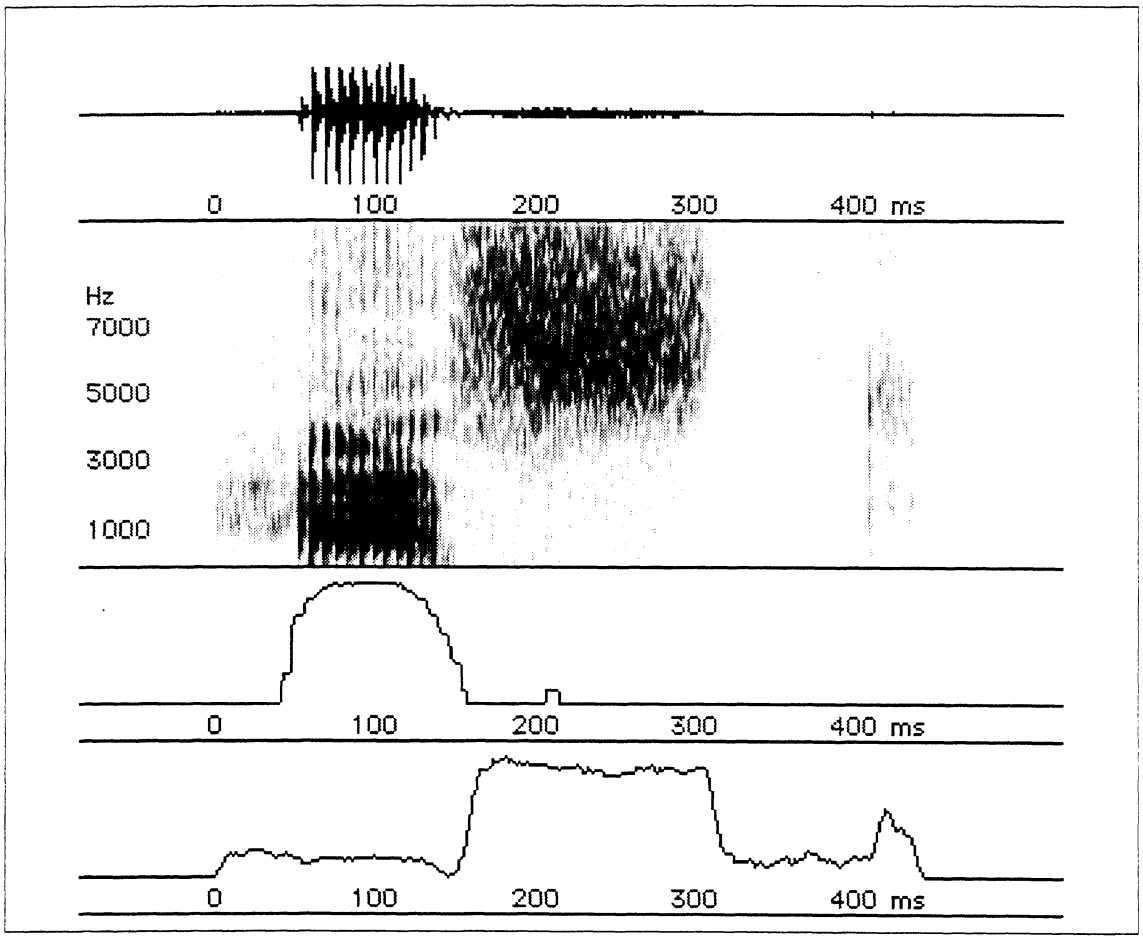

Figure 19: *<hast> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <hast> with an overall duration of 437 ms.*

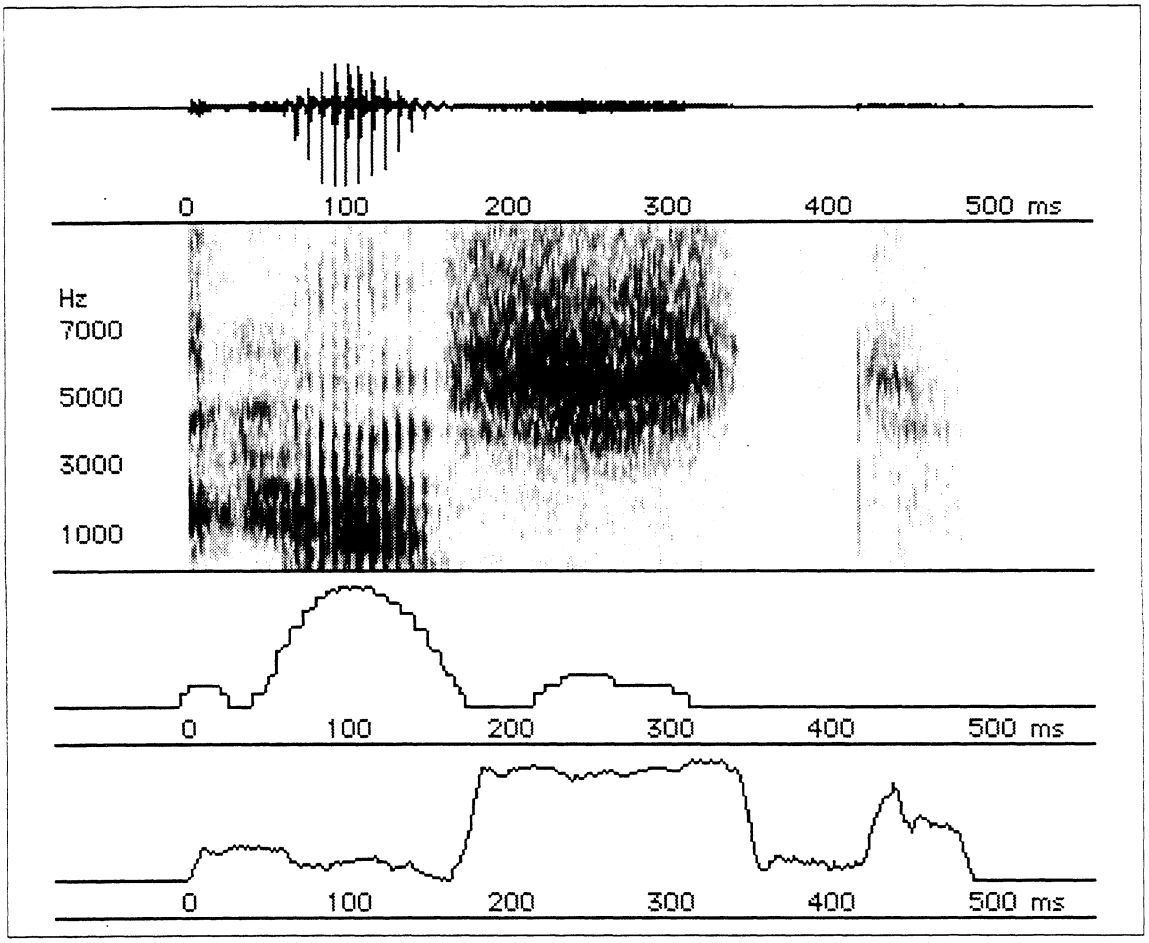

Figure 20: *<kast> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <kast> with an overall duration of 477 ms.*

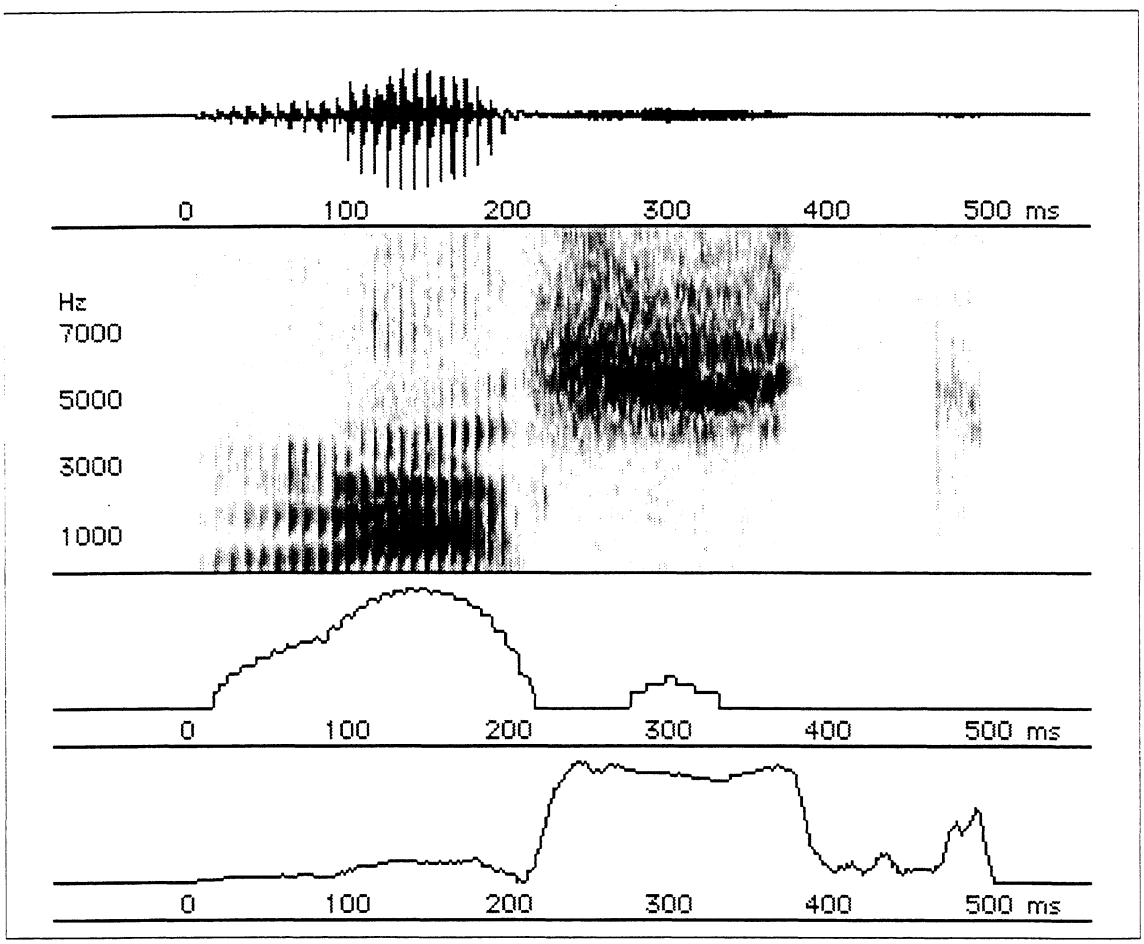

Figure 21: *<last> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <last> with an overall duration of 497 ms.*

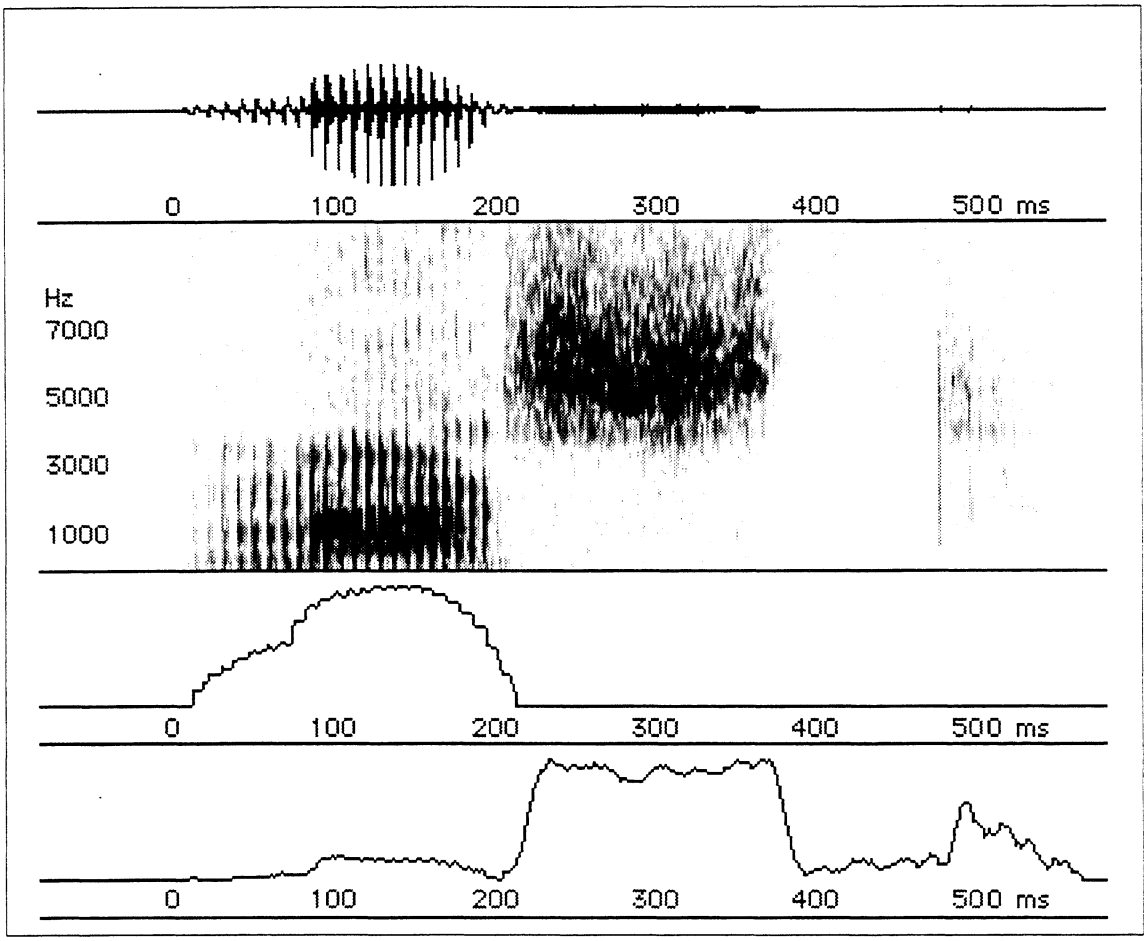

Figure 22: *<mast> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <mast> with an overall duration of 529 ms.*

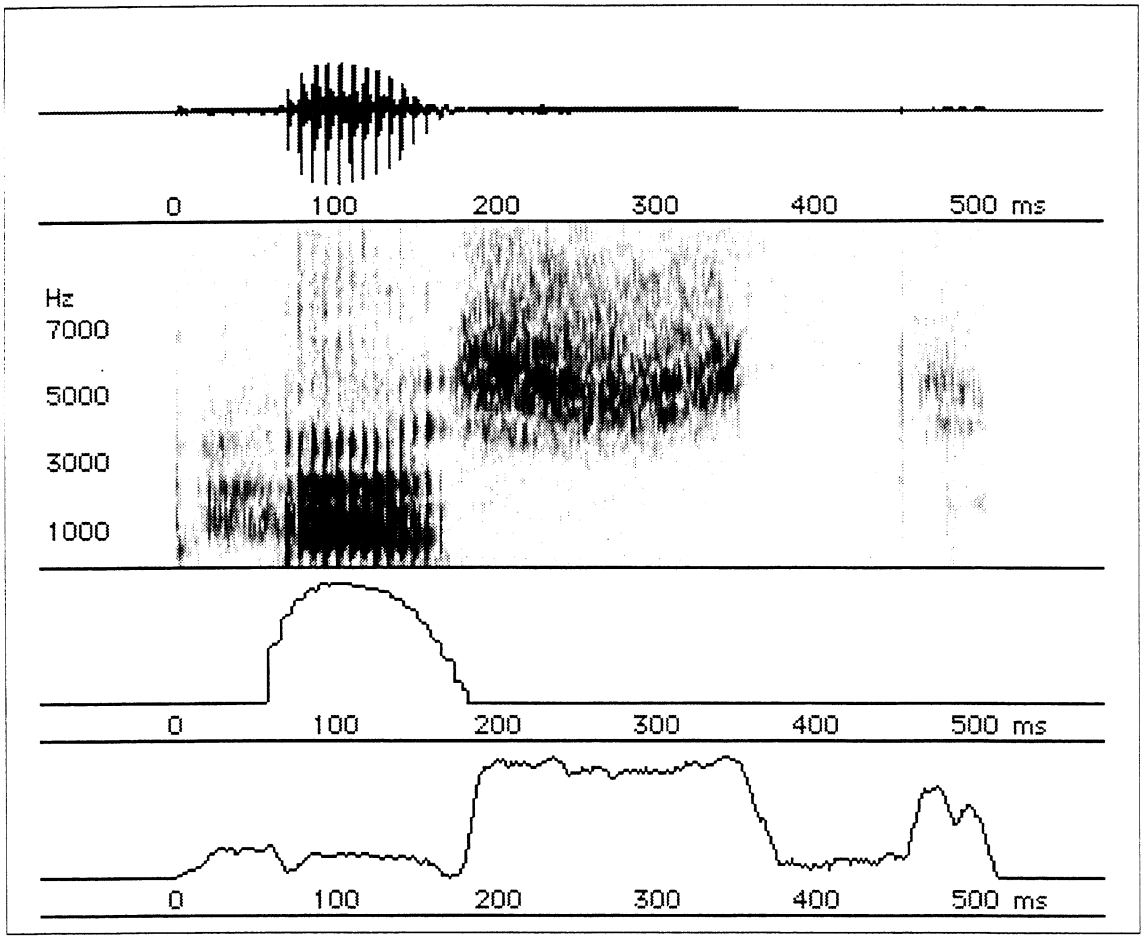

Figure 23: *<past> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <past> with an overall duration of 509 ms.*

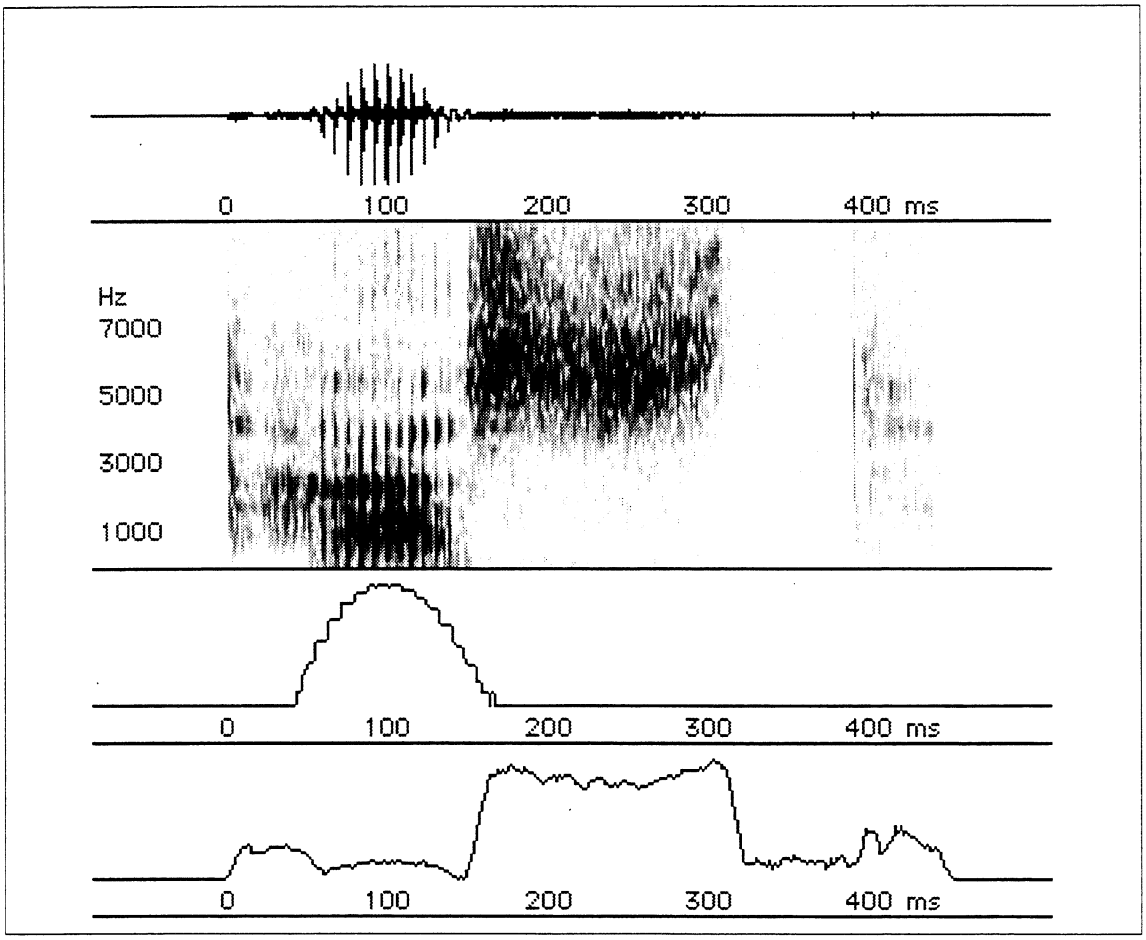

Figure 24: *<tast> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <tast> with an overall duration of 442 ms.*

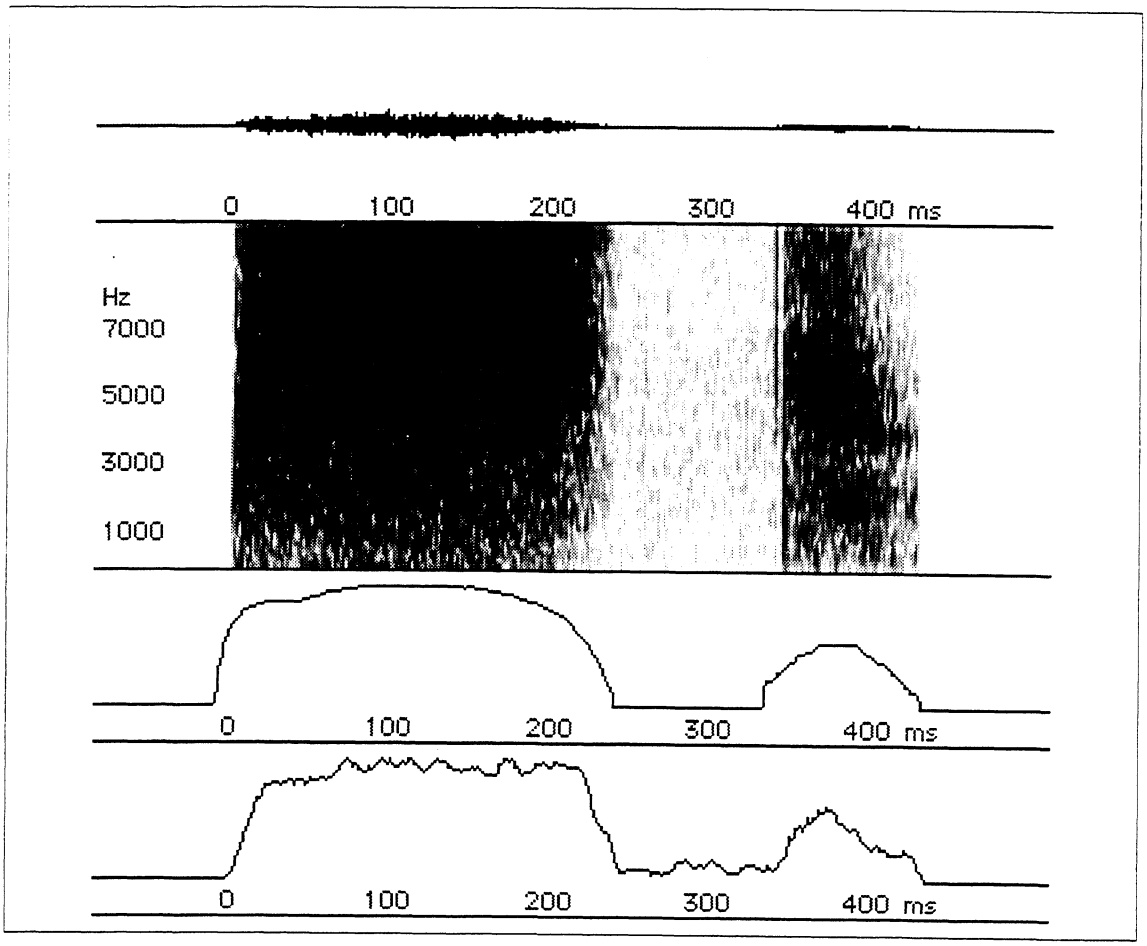

Figure 25: *<pst> Amplitude, sonagram, amplitude envelope (RMS, 30 ms) and zero crossing (10 ms) for stimulus <pst> with an overall duration of 428 ms.*