Face models based on a guided PCA of motion-capture data: Speaker dependent variability in /s/-/ʃ/ contrast production

Shinji Maeda
ENST/TSI & CNRS/UMR5141, Paris, France

We measure face deformations during speech production using a motion capture system, which provides 3D coordinate data of about 60 markers glued on the speaker's face. An arbitrary orthogonal factor analysis followed by a principal component analysis (together called a guided PCA) of the data has showed that the first 6 factors explain about 90% of the variance, for each of our 3 speakers. The 6 derived factors, therefore, allow us to efficiently analyze or to reconstruct with a reasonable accuracy the observed face deformations. Since these factors can be interpreted in articulatory terms, they can reveal underlying articulatory organizations. The comparison of lip gestures in terms of data derived factors suggests that these speakers differently maneuver the lips to achieve contrast between /s/ and /ʃ/. Such inter-speaker variability can occur because the acoustic contrast of these fricatives is shaped not only by the lip tube but also by cavities inside the mouth such as the sublingual cavity. In other words, these tube and cavity can acoustically compensate each other to produce their required acoustic properties.

1. Introduction

When data consist of a large number of variables having correlation structures between them, a factor analysis becomes effective. Motion capture data on the face deformations during speech production is such a case. In our experiment, a motion capture system measures 3D coordinates of individual markers glued on the speaker's face. Movements of markers are necessarily linked. The position of markers is affected by jaw gestures, by lip gestures like rounding and protrusion, and so on. Each of these gestures would deform the face in a particular way, creating a particular correlation pattern among markers' coordinates. Since the

measured face deformations present the sum of the effects of those gestures involved, markers' coordinates should have the correlation structure as the sum of individual correlation patterns.

Formally, let Y be a data matrix, which consists of observations of a set of variables. Y is centered and then normalized to obtain its Z score as

$$Z=(Y-m)/\sigma, \qquad (1)$$

where m and σ are, respectively, mean and standard deviation vector. A factor analysis only describes variations around the means of individual variables. Then, Z is assumed to be a weighted sum of factors as

$$Z=AX, \qquad (2)$$

where A is a matrix consisting of rows of weights, called factor patterns, specifying degrees of influence of corresponding factors upon individual variables. In other words, the sum of weighted factors presents the observed variations of individual variables. A factor analysis determines the factor pattern A and then values of factors X, *i.e.*, factor scores from A and Z.

The most basic factor analysis method is the principal component analysis (PCA) that determines factors so as to extract the maximum of variances. The derived factors, however, are not always interpretable. In comparison, an arbitrary orthogonal factor analysis (Overall, 1962) followed by PCA helps us to extract a set of interpretable factors from observed data (Badine *et al.*, 2002; Gabioud, 1994; Maeda, 1990). As in the case of PCA, the guided PCA derives an uncorrelated factor set. The total variance, therefore, becomes equal to the sum of variances explained by individual factors. If each of factor patterns can be interpreted in articulatory terms, the linear equation Eq. 2 can be considered as an articulatory model. In this report, we shall describe some examples for demonstrating the usefulness of such a data-derived functional model in analysis of face deformations during speech.

2. A face model based on a guided PCA of motion capture data

One American English male speaker (S1) and 2 French, male and female, speakers (respectively, S2 and S3) read a corpus consisting of a sequence of nonsense VCV syllables and a short text in the corresponding language. These 3 speakers were instructed to read the VCV syllables in a clear and hyper articulated way, and a text with 3 different speaking rates, slow, normal, and

rapid. For English, VCV sequences where V = /i/, /a/, or /u/ and C = one of 24 consonants are used. These 3 vowels and plus the high front rounded vowel /y/ are combined with 20 consonants in the French VCV sequences. The short English text consists of 28 syllables and the French text of 62 syllables.

Maeda et al. (2002) have reported, for S1, the details of the data acquisition and of the guided PCA analysis to extract uncorrelated articulatory factors that efficiently describe the measured face data. The same method was used for the data from the 2 French speakers. Briefly, a Vicon Motion Capture system with 6 infrared video cameras tracked 3D coordinates of 61 markers glued on the S1's face. For the 2 French speakers, 8 cameras tracked 63 face markers. For all the 3 speakers, common 61 face markers were approximately placed at the same relative locations. The camera speed was 120 frames/s for all the 3 speakers.

Before applying the guided PCA, effects of head movements on the position of markers are eliminated by head alignment. Y in Eq.1 becomes a matrix of head-aligned motion data of a speaker. For example, Y of S3 consists of 171 data variables, interlaced 3 coordinates of 57 markers in columns and 23164 frames of observations in rows.

First, we calculate the correlations between data variables (in Z), *i.e.*, a correlation matrix C. A factor analysis determines factor weights A. In the guided PCA, we specify factors to extract particular correlation structures. The first factor (f1), therefore, is determined so that it extracts the correlations between the z-coordinate of a marker on the chin, remaining 2 coordinates of the same marker, and 3 coordinates of all other face markers. Since this z-coordinate can be considered as a measure of the vertical jaw position, f1 represents the effects of vertical close/open jaw motions upon the face including the lips. Then the extracted correlation structure by f1 is subtracted from C. In this way, we determine, step-by-step the second (f2) and then the third factor (f3) representing, respectively, the effects of horizontal front/back jaw motions (along the y-axis) and those of horizontal left/right motions (along x-axis).

It may be interjected here that the skin, on which the chin marker is glued, can slide with up and down jaw motions, possibly causing a discrepancy between the measured marker movements and those of the lower jaw. As long as the marker movements are proportional to those of the jaw, the discrepancy would not affect our linear modeling. Since there is no guaranty about the proportional relationships, it would be more assuring to use a jaw splint (Badin *et al.*, 2002) fixed on the lower front incisors. With the Motion Capture system however, the use of such a device was not recommended, because it would disturb the real-

time automatic detection of markers' coordinates. We here assume therefore that measured chin marker motions are the reasonable representative of the jaw movements at least as a first-order approximation.

Second, the principal factors are determined from the residual of C after the subtractions of those first 3 jaw-related correlation structures. Table 1 summarizes variances explained by each of the first six factors determined for the 3 speakers.

Table 1: Explained variances (%) of the first 6 factors in a guided PCA. Factors in the columns are organized by functions and the corresponding factor numbers are indicated in the parentheses.

	Arbitrary orthogonal factors (Jaw gestures)			PCA (Intrinsic lip gestures)		
Speaker	high/low (f1)	front/back (f2)	left/right (f3)	round/ spread	protrude/ retract	(cheeks) lower/raise
S1 (English)	31	13	11	26 (f4)	7 (f5)	3 (f6)
S2 (French)	34	9	9	17 (f4)	3 (f6)	15 (f5)
S3 (French)	21	23	8	28 (f4)	3 (f6)	7 (f5)

In Table 1, the factors with number f5 and f6 are organized by functions, which we shall discus later. The first 6 factors explain about 90% of the variance for every speaker. Moreover, it is interesting to note that the cumulative variance over 3 jaw-related factors and that over the 3 lip-related factors vary little across speakers, respectively, about 53% and 37%. Nevertheless, there exist fairly important differences in the variance of individual factors across speakers. Note that S2 primarily uses the vertical jaw motion (f1 with 34% of variance) and much less front/back (f2 with 9%) and left/right motions (f3 with 9%). S3 uses more front/back (f2 with 23%) than high/low jaw motion (f1 with 21%). Moreover, S2 uses the cheek lowering/raising (f5 with 15%) to control the upper lip position, as explained latter. These findings suggest speakers employ different strategies in speech production. To make this point clearer, let us describe the effects of each factor upon the face deformation.

The effects of a factor can be visualized by varying the value of each factor from one extreme to the other while that of other factors is kept at zero. Figure 1 shows such visualization for selected factors for S1. The first 2 factors represent the effects of high/low (f1) and of front/back (f2) lower jaw motion in the order of their extractions, respectively in Figure 1a and 1b. The contribution of the vertical jaw motion is primarily lowering and elevation of the lower lip and to a lesser extent of the lip comissures. The center of the upper lip is hardly affected by the jaw motion. The other 2 speakers show similar patterns as the

consequence of jaw lowering and raising. The front/back jaw motion also primarily has an effect on the lower lip, which advances and retracts following jaw displacements. The Speaker S2 also exhibits this pattern, but not S3 as shown in Figure 2.

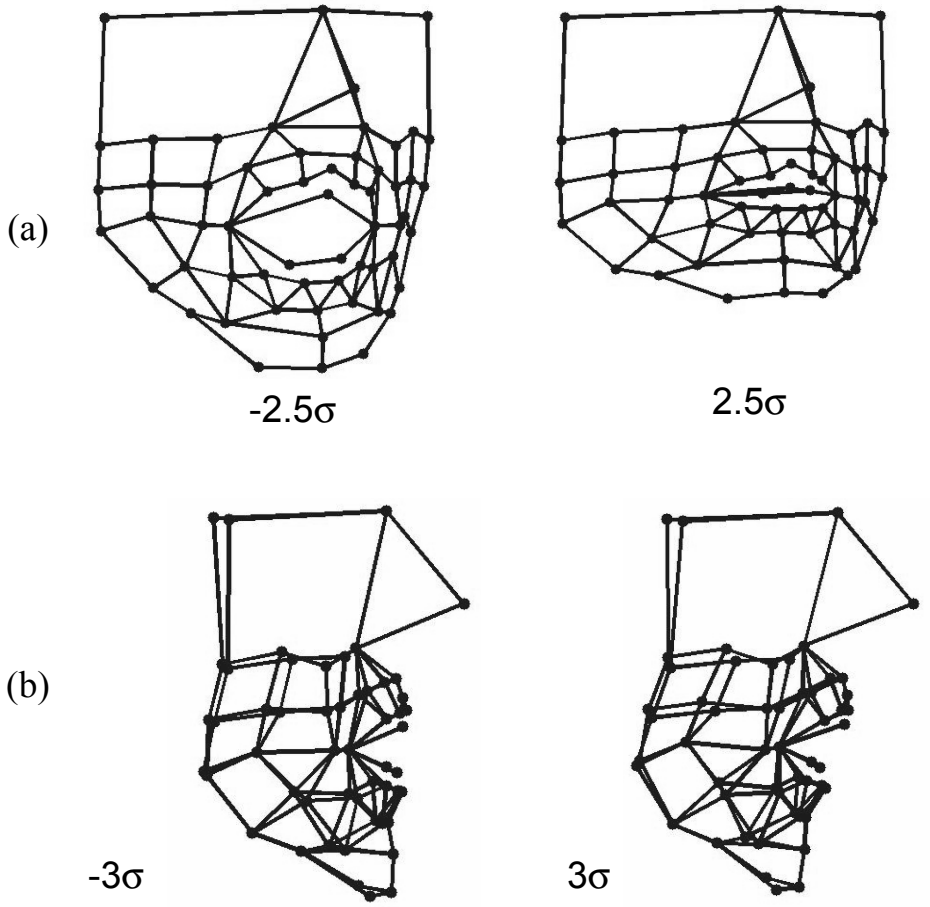


Figure 1: Model face of Speaker S1, where the value of each of 2 factors, f1 in (a) and f2 in (b), is varied from one extreme to the other as indicated while that of other factors is kept at zero. Faces in (a) and in (b) therefore indicate the effects of, respectively, jaw lowering/raising and front/back movements.

As mentioned before the speaker S3 is characterized by the high value of variance explaining the effects of the jaw front/back factor, f2 (23%), which is in fact greater than the variance of f1 (21%). Figure 2 shows that not only the lower lip deforms from front to back along the jaw movement, but also the upper lip appears to open up, resulting in a larger lip opening area in the back jaw position than in the front position. Presumably, the upper lip actively coordinates with front/back jaw movements, which is interpreted as a correlation structure in the factor analysis. This explains the high value of f2 variance observed in this speaker S3.

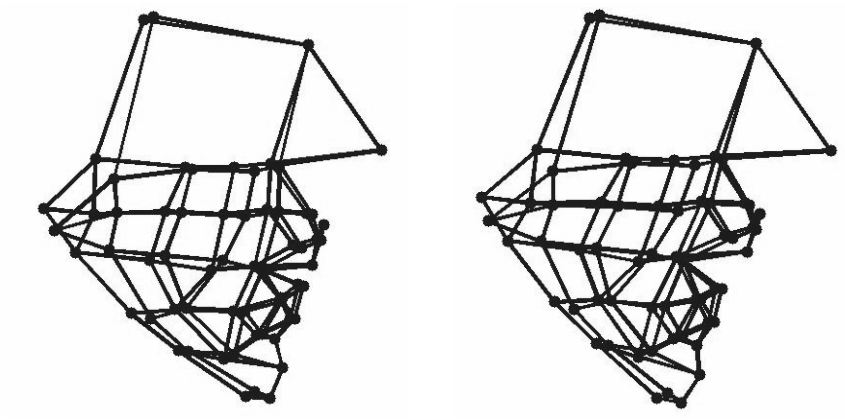


Figure 2: Model face of Speaker S3 where the value of the front/back jaw factor, f2, is varied from one extreme to the other as in Figure 1.

Badin et al. (2002) report a similar correlation between lip protrusion and jaw advance in their speaker. These authors consider the jaw advance as a part of lip related gesture and the lip-protrusion factor was extracted in the guided way, which resulted in a very high value of the explained variance. Here we assume a hierarchy in articulators. For example, the jaw gestures can be considered higher than the intrinsic lip gestures, since the influences of intrinsic lip gestures are localized in the lips themselves whereas that of the jaw extends over not only the lips but also the tongue and to an extent the larynx. In our analysis, therefore, the effects of the jaw positions were extracted before those of the lips following the hierarchical order.

The remaining 3 factors were determined by PCA and numbered in the order from high to low explained variance, as f4, f5, and f6. They must represent face deformations related not to the jaw gestures but to the intrinsic lip gestures, because they are determined on the residual of C after the subtractions of the correlation structures related to jaw motions in the 3 dimensions. The functions of these PCA derived factors must be visually identified by observing synthesized faces, as shown in Figure 3 for the speaker S1. It appears that the first PCA factor (f4) represents rounding/spreading in which the lip opening mainly deforms horizontally as seen in Figure 3a. The factor f4 of the French speakers, S2 and S4, also exhibits this kind of horizontal deformation of the lips. We interpret the factor f5 of S1 in Figure 3b as protrusion/retraction gestures involving both the lower and upper lips for this particular speaker. Although it may not be so visible in Figure 3b, the protrusion appears to be accompanied with a rotation of each lip to open up the aperture, which is visible in the video-clip. The factor f6 seems to represent a lowering/raising of the upper lip that is a direct consequence of the corresponding cheek movements, which is clearer in the f5 of S2, as seen in Figure 4.

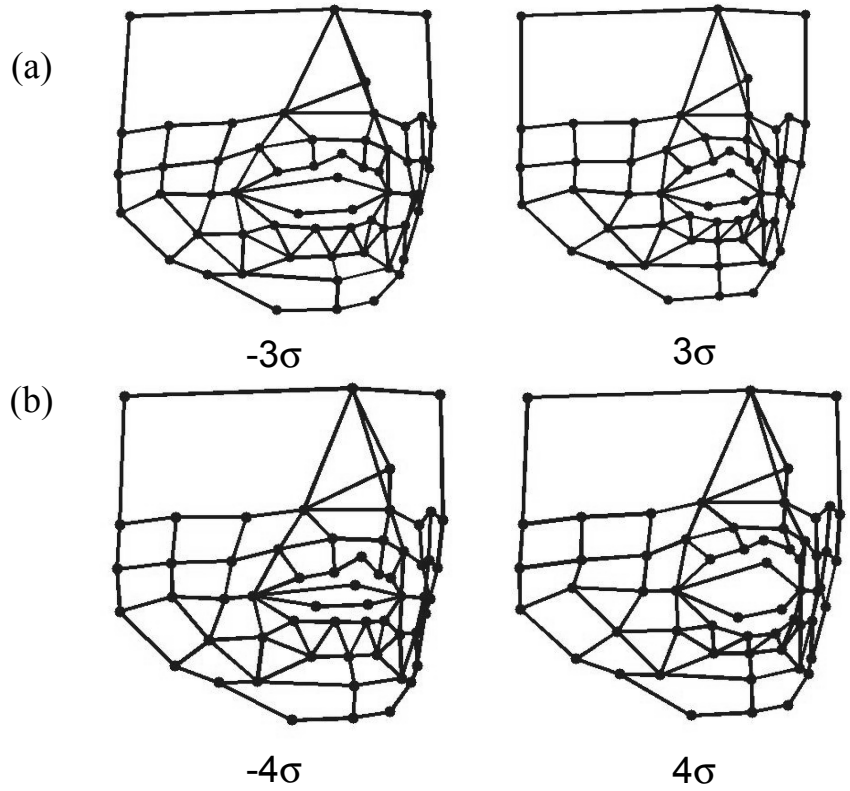


Figure 3: Model face of Speaker S1, where the value of each of 2 factors, f4 in (a) and f5 in (b), is varied from one extreme to the other as in Figure 1.

For the speaker S2, and also for S3, f5 seems to deform the face by lowering/rising of the cheeks as shown in Figure 4. The apparent larger lip opening is primarily due to a rising of the upper lip. In detail, the magnitude of cheek rise from the low position (with -4σ) to the high position (with 4σ) is greatest at the cheeks and then the upper lip, whereas the position of the lower lip is hardly affected. From this observation, we conclude that it's the cheeks which pull up the upper lip, and not the upper lip pushes up the cheeks. It may be obvious that the lips cannot push the cheeks up, if one considers the arrangements of the facial muscles (e.g., Gomi *et al*., 2002).

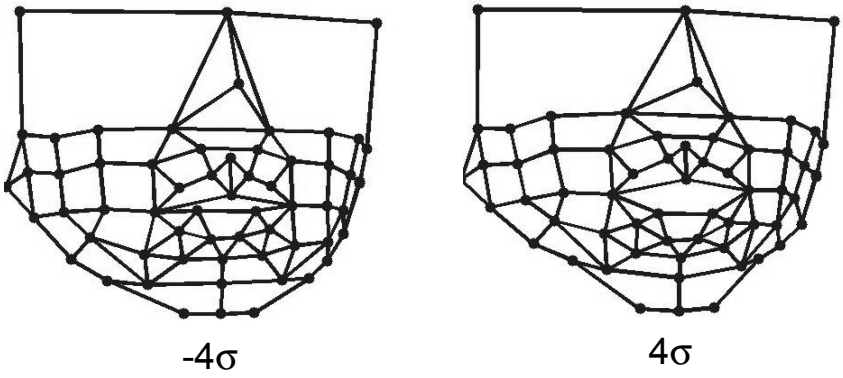


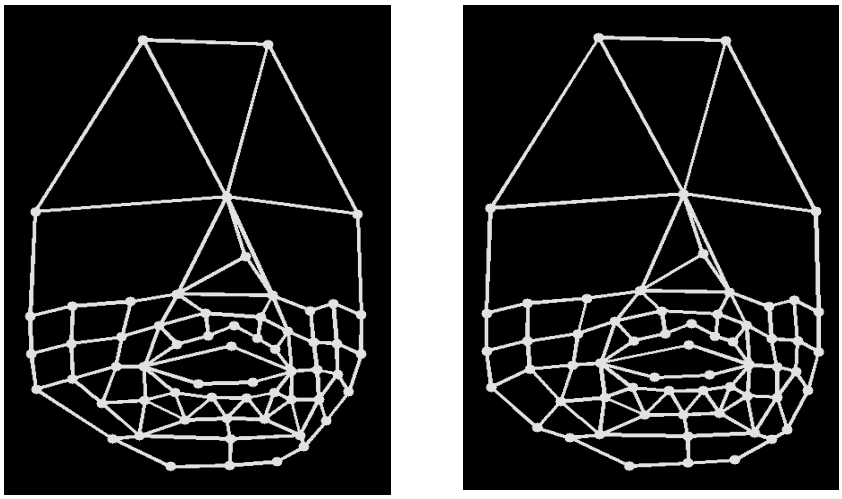
Figure 4: Model face of Speaker S2 where the value of the lower/rising cheek factors, f5, is varied from one extreme to the other as in Figure 1.

It may be noted here that functions of the 2 higher order factors, f5 and f6, are interchanged between S1 and French speakers S2 and S3, as already indicated in Table 1. For speaker S1, f5 represents the intrinsic lip gesture, retraction and protrusion and f6 having the smallest variance represents the cheek lowering and rise. For speakers S2 and S3, f5 accounts for the lowering/rising of the cheeks with the concomitant upper lip displacements and f6 that accounts for the lip protrusion has the smallest variance. It is safe to state therefore that S1 uses lip protrusion to control the length and aperture of the lip tube, while S1 and S2 employ the cheek lowering and rising to control the lip aperture. These observations suggest speakers use different articulatory maneuvers to produce speech sounds. In the next section, we shall describe in detail how differently speakers make the contrast between /s/ and /ʃ/ in terms of lip gestures.

Since each of those 6 factors can be considered as a functional elementary articulator, they tell us about how speakers articulate the lower jaw and the lips during speech production. As mentioned before, those first 6 factors explain about 90% of the variance for every speaker. Since the variance explained by any higher factor is less than 1.5 %, we discard factors higher than f6. In fact, it is not so evident to identify the functions of the higher factors because of their small individual influence on the face deformation. As a face model therefore, we use Eq. 2, but the full weight matrix A is replaced by its truncated version, A_6, for the first 6 factors, X_6, as follow:

$$Z=A_6X_6. \qquad (3)$$

Now, a synthetic factor model, Eq. 3, can be interpreted as an articulatory model as follows. The deformation Z from the mean face marker position is the sum of uncorrelated 6 linear components. Each weight, actually a set of coefficients of which number equals to that of variables (for example, 171 for S2 and S3), determines a particular face deformation pattern and the value of the factor, or the articulatory parameter, specifies the magnitude of that deformation. To obtain markers' positions in the original 3D coordinates, the deformation Z must be de-normalized using the inverse of Eq. 1. As described before, one of the interesting features of the face model is that the values of factors are calculated from the observed deformation Z and the truncated factor pattern A_6 in Eq. 3. Figure 5 shows the measured position of the markers during [i] and its reconstructed version with the first 6 factors. Note that markers' positions illustrated by dots are connected by lines so as to obtain a face like object. At present, we use this rather rudimentary face representation that was used already in Figures 1-5. The shapes of these 2 faces are hardly indistinguishable by eyes.



(a) (b)

Figure 5: The measured marker positions, indicated by the dots, of the speaker S1 during [i] in (a) and its reconstructed version using the first 6 factors shown in (b).

3. How speakers make [s] vs. [ʃ] contrast?

As described before, the values of the 6 factors can be directly calculated from the data and they should indicate an articulatory organization underlying the speech production. Let us show, as an example of a data analysis, how speakers make /s/-/ʃ/ contrast. It is known that /ʃ/ is produced with somewhat protruded open lips in English and French, whereas /s/ with unprotruded lips, although the lip shapes are phonologically unspecified (Gentil, 1980).

Figure 6 compares observed faces at about the center of [s] and of [ʃ] in /iCi/ syllables produced by those 3 speakers. All the lips appear spread due to the coarticulation of the vowel [i]; no matter the consonant is /s/ or /ʃ/. An obvious difference is that the lips are more open in /ʃ/ than in /s/ for all the speakers. Somewhat less obvious, but the lips appear to be protruded in /ʃ/, at least, in the speaker S1 and S2. These raw data clearly show the systematic geometrical differences in lip shapes, which are common to all the speakers. However, this doesn't necessarily mean that the same articulatory organizations underlie to produce the common geometrical differences, as describing in the next.

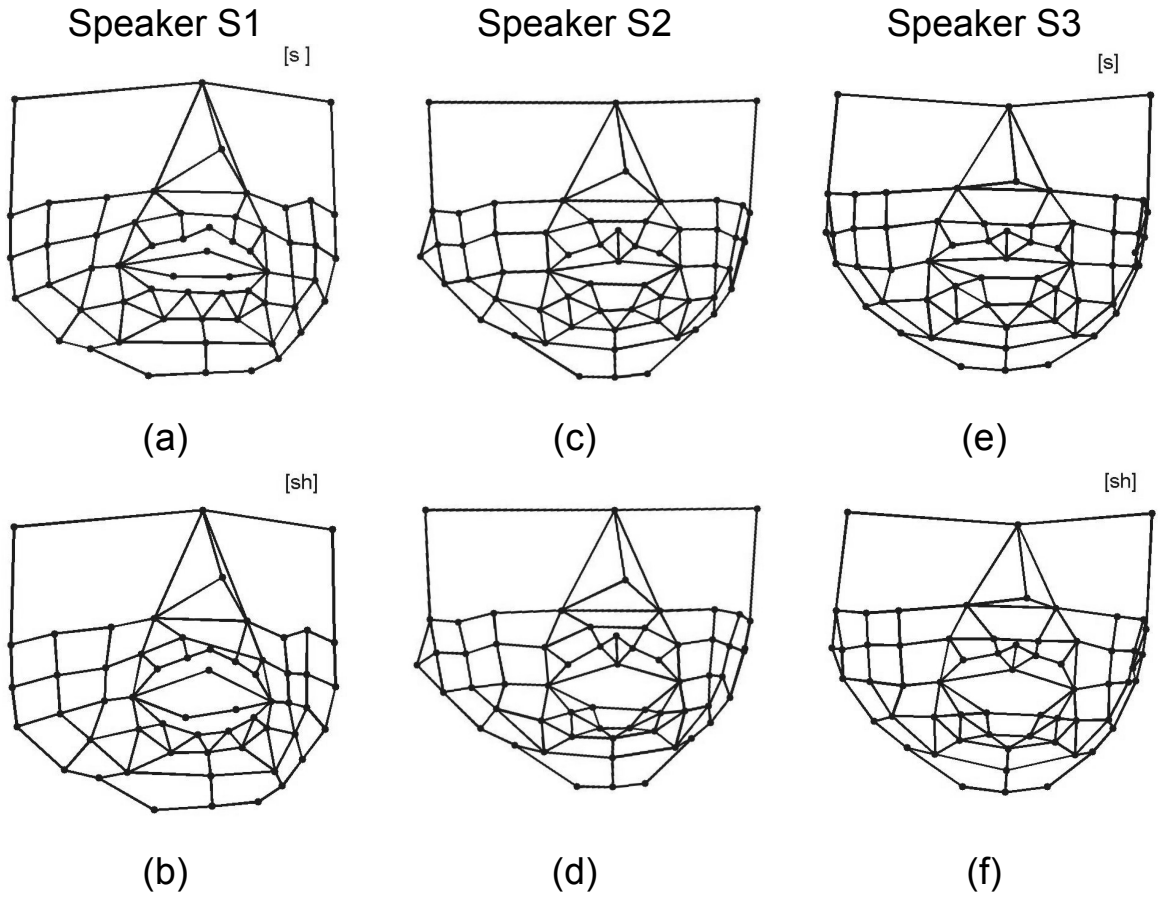
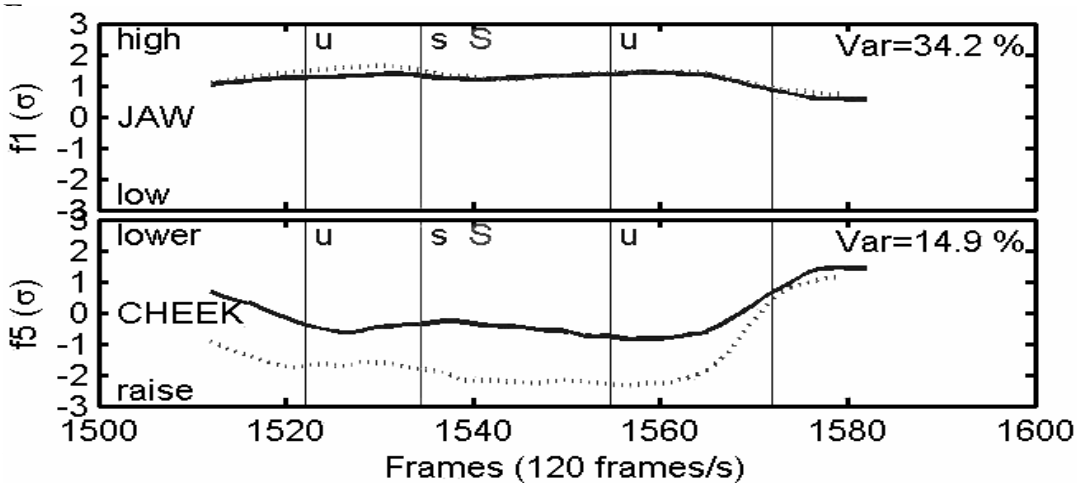


Figure 6: The measured faces at the center of [s] on the first row and of [ʃ] (denoted [sh] in the figures) on the second row. These plots are extracted from /iCi/ tokens uttered by the 3 speakers.

Figure 7 illustrates temporal variations of 2 selected factors, f1 and f6, along 2 syllables, [usu] in solid curves and [uʃu] in dotted curves, uttered by S2. Recall that f1 accounts for high/low jaw motions and f5 for the raising/lowering of the upper lip due to the cheek motions. In the case of f1, there is hardly any difference between the 2 curves, suggesting that f1 doesn't contribute to the [s] vs. [ʃ] distinction. The remaining factors also exhibit no important difference between [s] and [ʃ], except f5. In f5, shown in Figure 7, those 2 curves differ in an important way, suggesting the speaker S2 uses only f5 to differentiate [s] and [ʃ]. In other words, this factor (or this elementary articulator) is active for making the contrast. The speaker S2 raises the upper lip, which makes in turn the lip opening larger as seen in Figure 6d. The table 2 summarizes subjective judgments on which factors are active for speakers to make the contrast.



p

oral traces of 2 selected factors, f1 and f5, along syllables [usu] in solid curves and [uʃu] (denoted 'S' in the figures) in dotted curves, uttered by the speaker S2.

The "+" symbol in **Table 2** indicates the occurrence of marked difference between the values of a factor during [s] and the corresponding [ʃ] segment in a given vowel context. It is evident that 3 speakers use quite different articulatory organizations to produce the contrast. For example, S3 creates the contrast almost exclusively with the high/low jaw motion (f1). The use to the cheek raising/lowering, presumably to control the lip aperture, is unique to S2 among the speakers. Note that effects of vowel context are less pertinent than those of speaker difference. It may be noteworthy that the participation of cheek rise in S2 excludes with the vowel context /a/, suggesting an incompatibility of the upper lip rise in the open vowel context. Note also that no factor is active for making the contrast with the vowel context /u/ in S3. As reported by Gentil (1980), the anticipatory coarticulation of the labial attribute of the second vowel /u/ during /s/ could be the cause of the absence of the contrast. We shall give an additional discussion on this absence of difference in the lip gestures below.

Table 2: Active elementary articulators (factors) making /s/-/ʃ/ contrast

Factors		Speaker S1			Speaker S2				Speaker S3			
		aCa	iCi	uCu	aCa	iCi	uCu	yCy	aCa	iCi	uCu	yCy
JAW	high/low								+	+		+
	front/back			+	+	+						
LIPS	round/spread	+	+	+	+	+		+				
	protrude/retract	+	+	+								+
CHKS	lower/raise					+	+	+				

It becomes clear that speakers control the lip geometry to distinguish those 2 fricatives with quite different articulatory organizations. What acoustically accounts is the shape of the lip opening tube, which is roughly represented by its aperture (lip cross-sectional area) and its length. Although, our data using markers (i.e., flesh points) don't give us the exact geometry of the lip tube, it is safe to assume that the geometry is related to that of the flesh point representation, as seen in Figure 6, roughly in a proportional way. Then Figure 6 suggests that the lip aperture would be systematically greater in /ʃ/ than in /s/. The larger aperture of /ʃ/ than that of /s/ is created differently depending on speakers, by the combination of lip protrusion and spread in S1, by rising of the upper lip in S2, and merely by lowering of the jaw in S3. It isn't so systematic, however, in the case of the lip-tube length control. Speaker S1 and, a lesser extent, S2 seem to lengthen the lip tube with protrusion in /ʃ/ in comparison with /s/, which is neutral or slightly rounded. Note that protrusion/retraction in S2 is not marked in Table 2. This is because S2 protrudes the lips, to a certain degree, in both /s/ and /ʃ/. S3 does not lengthen at all in /ʃ/ relative to /s/.

Why is such an inter-speaker variation in lip geometry, especially lip tube length possible? In fact, Toda et al. (2002) have shown that differences in the observed lip configurations for /s/ and /ʃ/ alone cannot explain fairly important and distinctive differences in the spectral shape of these 2 classes of fricatives. The lower cutoff frequency in the noise spectrum of /s/ is much higher than that of /ʃ/ in the same vocalic contexts. Those authors have suggested that not only the differences in the lip geometry, but also those in tongue position and shape must contribute to the formation of the spectral characteristics of /s/ and of /ʃ/. In fact, Stevens (1993, 1999) has pointed out the acoustic influence of a relatively long sublingual cavity in /ʃ/ and the absence of such a cavity in /s/. Toda and Honda (2003) have confirmed Stevens' assertion by an MR imaging study. Then, the observed inter-speaker variations in the lip tube length can be explained as the consequence of an adjustment of the total length of the sublingual cavity plus the lip tube. As an extreme case, the lip gestures don’t differ much in the production of /s/ and /ʃ/ in the /uCu/ context (see Table 2) by the speaker S3. Presumably, this speaker makes the distinction only by tongue position. Extended studies on the labial and sublingual geometries in relation to their acoustic consequences are underway by those 2 authors and will be reported elsewhere.

4. Concluding remarks

We have shown in Section 2 that the guided PCA allows us to derive a compact and rational model of face deformations during speech. It is compact, because

the model consists of 6 uncorrelated (orthogonal) linear components. In other words, observed face deformations with apparent complexity have functionally, say, only 6 degrees of freedom. It is rational, because its 6 components can be interpreted by articulatory terms and thus, as shown in Section 3, the data analysis by factor values provides us with some insights about the underlying articulatory organizations.

Acknowledgements

The author is grateful to an anonymous reviewer and the editors Susanne Fuchs and Pascal Perrier for helpful feedback on an earlier draft. This work was supported, in part, by the project FEEDAT/ARC of the INRIA/LORRAINE, France.

References

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., and Savariaux, C., (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics*, 30, 533-553.

Gabioud, B., (1994). Articulatory models in speech synthesis. In *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art and Feature Challenges*, ed. E. Keller, John Wiley & Sons, 215-230.

Gentil, M. (1980). Sibilation et labialité en français – Coarticulation vocalique et valeur consonantique cible. *Labialité et Phonétique, Publications de l'Université des Langues et Lettres* de Grenoble, 181-201.

Gomi, H., Honda, M., Ito, T., and Murano, E., (2002). Compensatory articulation during bilabial fricative production by regulating muscle stiffness. *Journal of Phonetics*, 30, 261-279.

Maeda S., Toda M., Carlen A. J. and Meftahi L., (2002). Functional modeling of face movements during speech. *Proc. International Congress on Speech and Language Processing*, 17-20 September 2002, Denver (Colorado), 1529-1532.

Maeda, S. (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In *Speech production and modeling. (NATO Advanced Study Institute series)*, eds. W. J. Hardcastle and A. Marchal, Kluwer Academic Publishers, 131-149.

Overall, J. E., (1962). Orthogonal factors and uncorrelated factor scores. *Psychological Reports*, 10, 651-662.

Stevens K. N., (1993). Modelling affricate consonants. *Speech Communications*, 13, 33-43.

Stevens K. N., (1999). *Acoustic Phonetics*. The MIT Press.

Toda M., Maeda S., Carlen A. J. and Meftahi L., (2002). Lip gestures in English sibilants: Articulatory-acoustic relationship. *Proc. International Congress of Speech and Language Processing*, 17-20 September 2002, Denver (Colorado), 2165-2186.

Toda, M. and Honda, K., (2003). An MRI-based cross-linguistic study of sibilant fricatives. *Proceedings of the 6th International Seminar on Speech Production*, Sidney, December 7 to 10, 2003. 290-295.