Functional modeling of the face during speech production

Shinji MAEDA[1], Martine TODA[2], Andreas J. CARLEN[3], et Lyes MAFTAHI[3]

[1] ENST & URA820/CNRS, 46 rue Barrault 75634 Paris cedex 13
[2] ILPGA,Université Paris 3 & UMR7018/CNRS, 19, rue des Bernardins, 75005 Paris
[3] SyncMagic, 74bis, rue des Archives 75003 Paris
Email: maeda@tsi.enst.fr

ABSTRACT

We describe a functional modeling of face movements during speech. The data consist of face marker positions in 3D coordinates measured while a speaker read a corpus. An arbitrary orthogonal factor analysis followed by a principal component analysis on the data resulted in a set of five interpretable factors that explains 87% of variance. The first factor that account for the vertical jaw motion dominate open/close movement of the lips. Two principal factors describe, in out interpretation, the two intrinsic lip gestures, one specifies horizontal dimension, spread vs. round, and the other vertical dimension, open with rotation vs. close. Both the (horizontal) rounding and (vertical) opening contribute to the lip protrusion, which appear plausible from a biomechanical point of view.

1. INTRODUCTION

Modeling face movements (more precisely, deformations) becomes an important endeavor with increasing interest in audiovisual speech communications, in animations, and in speech training (e.g., [Bad00] and [Kro02]). The face modeling is also a natural extension of articulatory modeling, since articulators such as the lower jaw and lips determine face movements during speech.

We use a factor analysis on data to derive a model. The factor model describes observed face deformations as a linear combination of factors that could be regarded as the individual causes of the deformations, as jaw motion and particular lip gestures. Since such motions must be specific to a task, for example, mastication or speech, the factor model is not only descriptive but also functional. In addition to its linear nature, this is a limitation of any factor model in comparison with a model faithfully formulated on the biomechanics of the face.

It is known that any skilled movement, as speech gestures, would involve the coordination of a number of muscles. In such coordination, a group of muscle would work as a functional unit. An appropriate factor analysis can recover these functional units as the causes of face deformations. Although factor analysis doesn't tell us anything about the coordination itself, it is useful in several points. For example, it is often the case that a small number of factors can describe with a good accuracy the data having a large number of variables. The value of factors (factor scores) can be calculated from data in simple and straightforward way.

The factor representation, therefore, permits us to examine the movements with the small number of face parameters instead of the large number of raw variables or some selected raw variables. Moreover, if each factor reflects the coordinative unit, we may have a better chance to uncover phonetic/phonological rules of movements because the coordination must be purposeful and intended one to encode phonological message into the acoustic signal of speech.

In this paper, we focus, therefore, our attention to an elaboration of the analysis procedure for face movement data to obtain a set factors which are interpretable in biomechanical terms. Toda *at al*. report the phonetic analysis of the corpus using derived factors in another paper in JEP2002.

2. METHOD

2.1 Corpus

The corpus consists of 79 logatomes with type VCV and some VC, where V={/i/, /a/, and /u/} and C=24 English consonants, and 20 vowels, diphthongs, and retroflex vowels in /hVd/. The corpus also includes a short paragraph uttered with three different rates, fast, normal and slow. A male American subject read the corpus.

2.2 Data acquisition

Sixty-five reflective makers were glued on the subject face as shown in Figure 1. A Motion-Capture system tracked those markers. Six infrared video cameras captured images of the subject's face from different angles with the rate of 120 frames/s. A software then calculates 3D coordinates of markers' position. The position accuracy is claimed to be 0.1 mm.

It is noted here that what we called face movements is actually frame-by-frame variations of these marker coordinates.

2.2 Head alignment

Sequences of frames along logatomes and text are segmented into phonemes and labeled by the visual inspection of spectrograms of the simultaneously recorded speech signals. In order to remove the effects of head movements, face-marker coordinates of each frame are aligned by rotations and translations so that selected markers in a frame optimally match with those in the

arbitrarily chosen reference frame. The five selected markers are located on the forehead, labeled as A1 – A5 in Figure 1, and another on the nose, B5. Frame-by-frame plot of individual coordinates of a maker position exhibited small and rapid fluctuations and spike noises. The spike noises was due to the fact that the coordinates of a head marker(s) is sometime not detected and thus the number of markers for the alignment varies along a sequence of frames. The noises were removed by a low-pass filtering with the cutoff frequency of 10 Hz without noticeable deformation of the original marker movements.

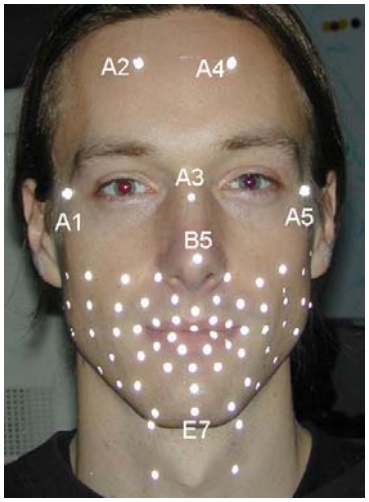

Figure 1:

Photograph of the subject with 65 face markers of which positions in the 3D coordinates are measured with the rate of 120 frames/s. Marker labels are added a posteriori for the description.

2.2 Preliminary factor analyses

Excluding four makers on the neck and six markers on the forehead and nose, 3D positions of 55 makers (165 coordinate values) that cover the lower part of the face are analyzed. Excluding also frames where more than five markers were not detected, 9014 frames constitute the data corresponding to 75 s of speech.

In preliminary experiments, we conducted the classical principal component analysis (PCA). The first four principal factors explain about 85% of variance, indicating a small number of functional dimensions of face movements. In detail, the first component, which explains nearly 50% of variance, seemed to account for the effect of lower-jaw vertical motion. The higher order components, however, were not clearly interpretable in phonetic/articulatory terms. In hope of obtaining a set of interpretable factors, face movements determined by the first four factors, truncating factors higher than five, were reanalyzed using an independent component analysis (ICA), specifically with JADE algorithm. The result was worse than that from PCA because the ICA tends to spread the variance more uniformly over the four factors, which makes interpretation even more difficult than before.

2.2 Arbitrary orthogonal factors and PCA

If we know the cause of the variance and one of observed variables can be assumed to the measure of that cause, there is a way to "transform" the variable into one of uncorrelated factors, which is called an arbitrary orthogonal factor. The correlations explained by the arbitrary factor are subtracted from the original correlations among variables. The residual correlations can be subjected to PCA. The method described in [Ove62] guarantees that the arbitrary factor and follow-up principal factors are orthogonal (or uncorrelated) to each other. If more than one variable could be regarded as measures of different causes, the corresponding arbitrary factors can be determined one-by-one, by subtracting the correlations explained by the factor at each step.

The marker E7 as measure of lower-jaw position: This method was effective in the analysis of midsagittal x-ray images of the vocal tract during speech. A lower jaw position measured on the images was assumed to be one of causes of the vocal-tract deformations. Only in this way, intrinsic tongue deformations could clearly dissociated from the effects of jaw movements [Mae78].

The current face movement data lack jaw position measure however. An alternative is to use a marker on the chin, as the marker E7 (see Figure 1), as a measure of jaw motion. This faces a common objection that the skin on the chin can slide relative to the moving jaw and so as the marker glued on the skin surface. The maker, therefore, doesn't exactly follow the jaw movements. We decided to investigate whether this is well being the case for our speaker.

If markers on the chin faithfully followed the jaw motion, the movements of a marker would be completely predicted from any one of the markers. To test this hypothesis, we selected the three center markers on the chin, E7 and other two markers just above E7, E6 and M11 (not shown in Figure 1). The three coordinates of E7 are assumed as arbitrary factors describing jaw position. Then the arbitrary orthogonal factor analysis calculates the residual of the global variance of the markers that cannot be explained by the E7 movements in the 3D space. The calculation shows that the residual is only 3.6%. We have judged, at least for the current speaker, that E7 can be regarded as a reasonable measure of the lower-jaw motion.

The next question is whether we need all the three coordinates to describe the jaw motion. Dispersion of the dark points in Figure 2 indicates measured positions of the marker E7 in the 3D space. Gray points indicate its 2D projections onto the x-y (coronal), y-z (sagittal), and x-z (frontal) planes.

As evident in Figure 2, the marker movements are the largest in the vertical high/low dimension (σ=4.9mm) and then in the horizontal front/back dimension (σ =3.2mm). These two movements appears to correlate each other as suggested by the projected points distributed along the principal axis on the y-z plane. Presumably, the principal axis corresponds to the open/close motion of the lower

jaw. Note that there are considerable left-right movements as seen in the projection onto the x-y plane (σ =1.5mm).

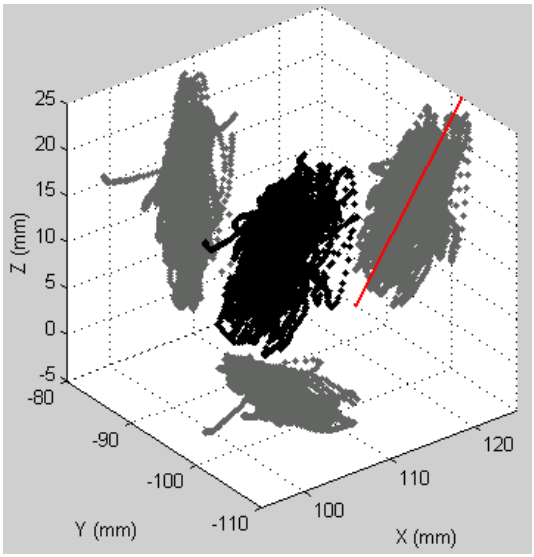

Figure 2. Observed position of the marker E7 in the 3D coordinates indicated by dark points. Dispersions of gray points illustrate the projections. The x-axis corresponds to left/right, the y-axis to the front/back, and the z-axis to high/low dimension. The solid-line on the y-z plane is the principal axis.

Since the position of E7 spread in the 3 coordinates, there is no justification to eliminate any E7 dimension from the consideration. Moreover, the factor analysis of E7 movements in the three dimensions indicated that 18.6% of variance remains as the residual for the front/back displacements, in spite of its apparent high correlation with the dominant vertical movements.

Factor analysis of face movement data: In the factor analysis of the face data therefore, we first removed the influences of jaw motion in the three dimensions upon the face deformations by the three arbitrary factors. The order of extraction was the high/low movements in the z-coordinate, front/back in the y-coordinate, and then left/right in the x-coordinate. Second, principal factors are determined from the residual correlations.

Table 1: Proportion of variances explained by the first five factors in % (on the fourth row) and the cumulative variances in % (on the fifth row).

Arbitrary factors (JAW)			PCA	
high/low	front/back	left/right	first	second
f1	f2	f3	f4	f5
31.4	12.9	10.5	25.8	6.7
31.4	44.3	54.8	80.6	87.3

3. RESULTS

The variances explained by the first five factors (three arbitrary factors that account for jaw motion, and two principal factors) and the cumulative variances are shown in Table 1. The five factors together explain 87.3% of the variance and the principal factors higher than third were truncated.

Effects of three arbitrary factors: Figure 3 visualizes the effects of the three jaw factors, i.e., f1, f2, and f3 in Table 1. Thin lines connect the markers at their average (or a rest) position so that together they resemble a face. The thick dark and thick gray vectors corresponding, respectively, to a positive and a negative value of the factor depict displacements in opposite direction of each marker.

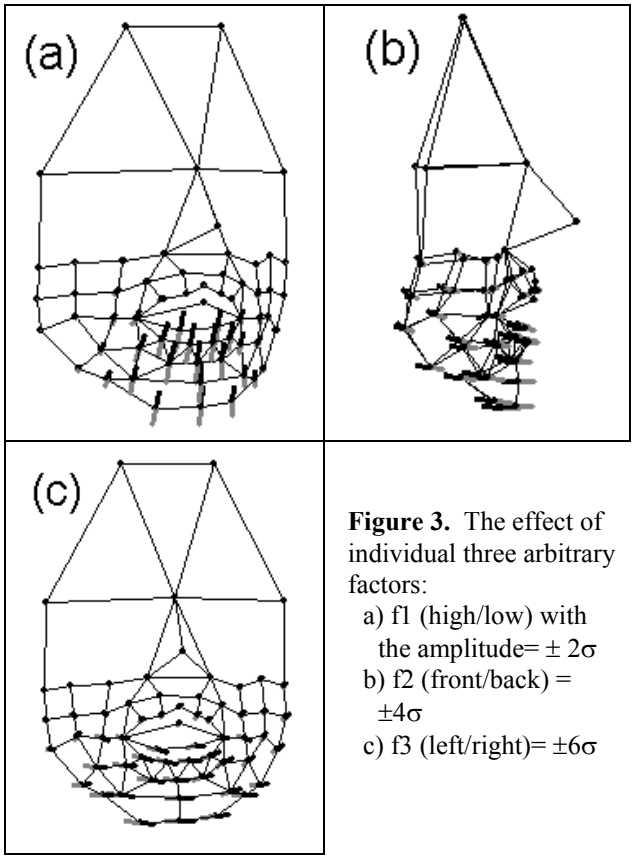

Figure 3. The effect of individual three arbitrary factors:
a) f1 (high/low) with the amplitude= $\pm 2\sigma$
b) f2 (front/back) = $\pm 4\sigma$
c) f3 (left/right)= $\pm 6\sigma$

Except the front/back jaw movements in Figure 3b, the effects are localized to the chin region, in particular for the high/low movements in (a). Because of this concentration of the area, the effect of high/low jaw movement on the mouth is much more than the value of the explained variance 31.4, would suggest. Recall that the explained variances are normalized with the total variance of all the 55 markers. This factor f1, therefore, plays the major role in close/open of the mouth during speech.

The left/right movement is well present with the explained variance of 10.5%. A faithful reproduction of our subject during speech certainly requires a correct specification of this left/right movement. We are not so certain that such an accurate specification is needed for the acoustics of the lips and thus for speech synthesis. Moreover, the upper lip is hardly affected by the jaw movements, which confirms common observation.

Effect of two principal factors: The visualization of the effect of the first and second principal factors, respectively f4 and f5, are depicted in Figure 4. Since these two factors are derived from the PCA on the

residual correlations where all the influence of jaw motion is removed, we must interpret the results in terms of functional or biomechanical significance. The visualization helps the interpretation.

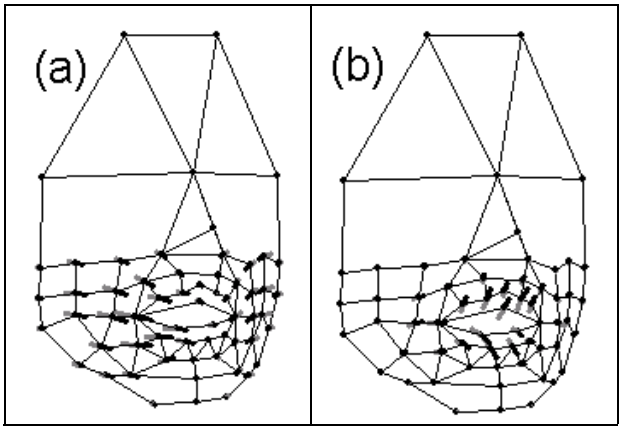

Figure 4. Effect of the first principal factor (f4=±6σ) in (a) and second factor (f5=±10σ) in (b). They can be interpreted as horizontal (round/spread) and vertical (open/close) lip gestures, respectively.

The marker movement vectors of the two factors exhibit a horizontal symmetry, left/right for f4 and a vertical symmetry, above/below for f5, with respect to the center of the lip opening. We take the lip-centered dispersion of vectors as the manifestation of lip gestures. Moreover these symmetries suggest that symmetrically opposing pairs of the muscles surrounding the lips are activated in coordination.

During speech, therefore, the lips can be controlled in the two orthogonal directions, horizontal and vertical. For the factor f4, its negative value corresponds to a spreading of the lips, presumably the activation of the risorius and other extrinsic muscles. Notice that the movement vectors cover the cheeks and their direction appears along the face surface. The vector amplitude appears to considerably diminish near the midsagittal line of the lips in Figure 4a. The apparent diminishment is in part due to a change in the direction of the movement. At the lip center the orientation of the vector is perpendicular to the face, i.e., in the front/back directions. As the factor value becomes negative therefore, the lips not only deformed to the opposite of the spread configuration, i.e., a horizontal closure, but also protrude, resulting is a lip rounding. Presumably, this rounding gesture is realized by the simultaneous contraction of the orbicularis oris superior and inferior. Note then that the effects of the factor f4 appear to account for functionally antagonistic activation of the orbicularis oris muscles and the risorius in the rounding and spreading of the lips.

The deformations due to the factor f5, shown in Figure4b, is localized within the lip region, suggesting its relation to the activity of the orbicularis oris superior and inferior, and perhaps some nearby muscles such as the depressor labii inferioris and levator labii superioris. The lips vertically open or close as the factor value varies between positive and negative values, correspondingly. Moreover, the open/close movements accompany forward/backward motion with a concomitant rotation. As the consequence, a rotational protrusion of the lips would occur with the opening, which is a common observation in the production of French rounded vowels. [Hon95] has postulated that the contraction of peripheral layers of the orbicularis oris is the underlying mechanism of this rotational protrusion. Note, of course, that a combination of these two intrinsic lip factors and in addition the three jaw arbitrary factors determines the shapes of the lips.

4. CONCLUDING REMARKS

The arbitrary orthogonal factor followed by PCA appears effective to formulate a functional face model. Its compact description of face deformations should be useful as a synthetic model, as well as a data transformation for phonetic/phonological analysis. This approach allows us to build a model without complications of often-unknown physical properties of tissues and muscles and their complex anatomy.

ACKNOWLEDGEMENT

We thank Rémi Brun, Attitude Studio (Paris) for his professional expertise in the data acquisition. This work is supported, in part, by the project SAALSA No. 7215/17/00 in the program PRIAMM of CNC.

BIBLIOGRAPHIE

[Bad00] Badin, P., Borel, P., Bailly, G., Revéret, L., & Baciu, M. (2000), "Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue lips and face based on MRI and video images", 5th Seminar on Speech Production: Model and Data, pp. 261-264

[Hon95] Honda, K., Kurita, T., Kakita, Y., & Maeda, S. (1995), " Physiology of the lips and modeling of lip gestures ", Journal of Phonetics, Vol. 23, pp. 243-254.

[Mae78] Maeda, S., (1978), " Un modèle articulatoire de la langue avec composantes linéaires ", 10émes JEP, GALF, pp. 152-164.

[Ove62] Overall, J.E., (1962), " Orthogonal factors and uncorrelated factor scores ", Psychological Reports, Vol. 10, 651-662.

[Kro02] Kroos, C., Kuratate, T., & Vatikiotis-Bateson, (2002), "Video-based face measurement", Journal of Phonetics (to be appeared)