

COMPENSATORY ARTICULATION IN SPEECH:
ANALYSIS OF X-RAY DATA WITH AN ARTICULATORY MODEL

Shinji Maeda

Département de Recherche en Communications par la Parole
Centre National d'Etudes des Télécommunications (CNET)
22301 Lannion, France

ABSTRACT

Roughly 1000 frames of cineradiographic and labiofilm data on the vocal tract corresponding to 10 French sentences uttered by two speakers have been analyzed statistically. The analysis resulted in an articulatory model consists of a limited number of linear components. With this model the temporal variations of the vocal-tract shapes are described by the frame-to-frame samples of the articulatory parameters. We observe that "target" parameter values for the same vowel vary significantly presumably due to different phonetic contexts. An acoustic calculation with the model predicts that a particular pair of articulators can compensate acoustically each other. For example, by an appropriate adjustment of the tongue-dorsal position, the model is capable of producing the same F1-F2 pattern for different jaw position, or *vice-versa*. The compensation between the jaw and the dorsal positions, however, is possible only for unrounded vowels. In the case of rounded vowels, the jaw position can be compensated by the lip aperture. The measured "target" values of the paired parameters indicated a linear relationship, suggesting that the speakers actually exploit the inter-articulator compensation in the speech production. This explains the observed large "target" value variability. The comparison of parameter trajectories for the same sentences uttered by the two speakers indicates more similitude than difference, suggesting that the manner of the production involving the compensatory articulation could be relatively invariant.

1. INTRODUCTION

In continuous speech, we observe that the acoustic manifestation of each phoneme deviates considerably from its intrinsic characteristics, and depends on the phonetic context. Ohman (1966) has demonstrated that the formant trajectories in VCV (vowel-consonant-vowel) syllables are determined by a vowel-to-vowel gesture plus the consonant articulation. The acoustics of the consonant is, therefore, influenced by the surrounding vowels. The coarticulatory "model" implicitly assumes a fixed vowel target position of each articulator for a given vowel. The vowel-to-vowel gesture is specified as the displacement of each articulator from one vowel target to another. A local consonantal articulation is superimposed on this vowel gesture. What is crucial here is the validity of a set of "fixed" articulatory targets which provides an anchor point and therefore makes the description of the coarticulation simple and explicit.

There are, however, numerous counter indications against fixed articulatory targets for vowels or in general for phonemes. The acoustics of a vowel can be modified as a function of the articulatory place of the following consonant, *e.g.*, the palatalization or velarization of the vowel. When the intervocalic consonant is a labial, the preceding vowel can be coarticulated heavily with the second vowel to such extent that its formant trajectories of the entire portion are modified (Vaissière, 1987). The degree of the coarticulation tends to increase with speaking rate.

Lindblom *et al.* (1971) have demonstrated the existence of jaw-tongue compensation from their bite-block experiments. On one hand, they showed that a speaker can produce almost identical acoustic patterns of vowels regardless of the fact that the jaw position is fixed by the bite-block or not (also in Lindblom *et al.*, 1979; Gay *et al.*, 1981). On the other hand, a model analysis of

x-ray data on various sustained vowel articulations indicated an ordered relation between the jaw position and the vowel height in the phonological sense. Without some constraints such ordered relation shouldn't exist, since a vowel can be produced with almost any arbitrary jaw position. In order to account for the observed ordered jaw position, they have postulated a principle of "minimum antagonism". In the articulation of a vowel with a given height, there is an optimum jaw position so that the antagonism between the jaw and the tongue-body gesture becomes minimum. The corollary is that the optimum jaw position is proportional to the vowel height. However its validity could be limited to the case of vowels in isolation. If such a principle were generalized literally to arbitrary sentences, then we would expect every vowel to be produced with their locally optimum and thus fixed targets, which is unlikely the case. Nevertheless, it is interesting to assess the presence or absence of fixed articulatory targets for vowels in the utterance of sentences.

In this paper, we shall document data in favor of the premise of the compensatory articulation operating in the production of vowels in "natural" speech (without a bite-block), and against the fixed target hypothesis. A salient feature of this study is that the simultaneous cineradiographic and labiofilm data are analyzed by using an articulatory model. The observed time-varying vocal-tract shapes are described in terms of frame-by-frame variations of model parameters, such as the jaw and tongue-dorsal positions *etc.*. Since the model specifies the entire tract shapes, the acoustic characteristics are calculated for a given set of parameter values. This is important in that it allows us to describe and evaluate the consequences of a change in the individual articulator positions in terms of the acoustics, for example, of formant patterns.

2. ARTICULATORY DATA

The data consist of more than 1000 digitized tracings of vocal tract shapes corresponding to 10 French sentences uttered by two female speakers, PB and DF. Each of the data frames representing the vocal tract profiles, from the glottis to the lip opening, and the frontal lip shapes was obtained by manually tracing the simultaneous radiofilm and labiofilm with 50 frames per second (see Fig.1a and Fig.1b). Detailed information, such as the list of the sentences, selected x-ray tracings and the corresponding acoustic data, is found in Bothorel *et al.* (1986).

3. MEASUREMENT OF VOCAL-TRACT SHAPES

The first step of the data analysis is to measure the vocal-tract shapes with appropriate references. This procedure is necessary because there is no frame-to-frame correspondence in a series of points which constitutes a particular contour representing, for example, the midsagittal tongue shapes. The vocal tract is divided into three sections, the lip-opening tube, the principal tract corresponding to the buccal and pharynx cavities, and the laryngeal extreme. The mandibular position is also measured for each frame.

The form of the interior and exterior tract walls in the principal section is measured using a semipolar coordinate system as shown in Fig.1a. The coordinate is fixed to the rigid maxillary structure. All movements, or deformations of the tongue, of the velum, of the rear pharynx walls, and of the larynx are therefore measured with relative to the maxilla. The intersections between the contours and the grids, indicated by the dots in Fig.1a, are detected by an automatic procedure based on a binary search algorithm.

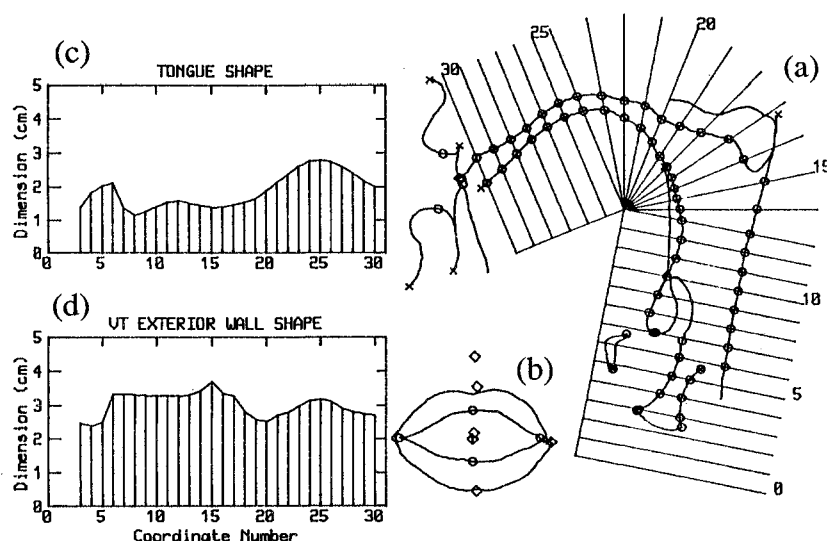


Figure 1. Measurements of the vocal-tract shapes. The dots indicate the measured points. (a) Profile superimposed with a semipolar coordinate system. (b) Frontal view of the lip-opening. (c) A vector representation of the measured tongue shape including the upper laryngeal region. (d) A vector representation of the exterior walls.

A set of the coordinate values of the intersections in the order of the grid number then constitutes a vector representing the measured interior contour as shown in Fig.1c and exterior contour in Fig.1d. The vectors are considered as the tract shape sampled at the coordinate grids.

In the case of the lip and the larynx tubes, the semipolar coordinate is useless, since these articulators tend to move in the direction perpendicular to the grid lines. Each of the two points corresponding to the anterior and the posterior extremes of the larynx edge is specified by two values (x, y). The form of the larynx tube is specified by combining these two points and the intersections detected in the previous procedure.

The lip-opening tube is modelled by a uniform elliptic tube with three variables, height, frontal width, and length (protrusion). The height and width are defined as, respectively, the distance between the highest and lowest points, and the distance between the most left and right points on the front inner lip contours. The four detected points are indicated by dots circles in Fig.1b. The protrusion is measured on the lip profile as distance between the upper incisors and the point where the vertical separation of the upper and lower lips becomes minimal. These definitions of the three variables enable us to use a simple algorithm in the automated measurement.

Vocal-tract profiles can be regenerated by projecting the measured vectors, shown in Fig.1c and Fig.1d, on the reference coordinate system. Combining the profile with the lip tube and the larynx edge, the entire vocal tract shape is specified. The area function and then the transfer function and the formant frequency values are calculated from the regenerated vocal-tract. The calculated vocal-tract transfer is important in that we actually model the measured vectors, and not directly the original vocal-tract contours; therefore the transfer function can serve as a reference in the evaluation of the model in terms of its acoustic performance.

4. LINEAR COMPONENT MODEL OF THE TONGUE

An articulatory model is derived as the result of a factor analysis on the data. This does not necessarily mean that the model comes out automatically from such an analysis. We must have a basic idea about how to represent the vocal tract. In this regard, we follow a jaw based model proposed by Lindblom *et al.* (1971), in which vocal tract shapes are determined as a function of parameters as jaw, tongue-body, tongue-tip, lip height and width, and larynx height. The underlying hypothesis is that, during speech production, the complex activities of the articulatory organs are organized into a limited number of independently controllable functional blocks and that the tract shapes are determined by the state of these blocks. Let us call these blocks as "elementary articulators" and their action as "elementary gestures". We attempt to extract these elementary articulators from the data by means of a statistical analysis.

The tongue shape of each frame can be described by the measured vector with about 30 variables, such as shown in Fig.1c. Their values depend on the states of the individual elementary articulators. The basic assumption is that the influence of each elementary gesture on the deviation of the tongue from its neutral shape is proportional to a parameter representing a strength of that gesture, and that the influences from the different articulators can be added up to produce the final tongue shapes. In short, the measured vectors, thus the vocal-tract shapes, are represented by a linear combination of the effects of elementary gestures.

The measured vocal-tract data are subjected to the so-called arbitrary factor analysis (Overall, 1962; Maeda, 1979). The results indicated that the tongue contours can decompose into four articulatorily relevant components representing the effects of the jaw, tongue-dorsal position, the tongue shape, and tongue-tip position. The lip section is specified by the jaw position plus two parameters, the height and protrusion. The width is predicted from these three parameters. The larynx edge depends the jaw and its own height. The vocal tract shapes, therefore, are specified by the seven parameters. Because of the truncation of higher components, errors in the specification of the shapes occur, particularly in the laryngeal region. Consequently, the calculated transfer function exhibits a certain error at high frequencies. The current model, however, can be useful in investigating the articulatory-acoustic relationships in terms of, say, the first three formants.

5. ARTICULATORY-ACOUSTIC RELATIONSHIPS

An important advantage of an articulatory model specifying the entire vocal-tract is that the relationships between the articulation and its acoustic consequences can be studied independently of the data (Majid, 1982). In fact, such predicted relations help us to understand better the data representing the frame-by-frame samples of the individual articulatory gestures.

In Fig.2a, the acoustic effects of a change in the two articulatory parameters, jaw position marked by the open circles and tongue-dorsal position by the closed diamonds, are plotted on the F1-F2 plane. The effects are calculated around each of the five cardinal vowels. All parameters values, except the value of the jaw, or of the dorsal parameter, are set equal to those appropriate for each vowel. The parameter value in question is varied from -2.0 to 2.0 in seven steps. The magnitude of the variation is well within that observed during speech, since it rarely exceeds below -3.0 or above 3.0 for any parameter. The parameter values are normalized by the corresponding standard deviations.

The acoustic consequences of the two articulator movements are distinctively different for the two vowel groups, rounded vowels [u and o] and unrounded vowels [i, e, and a]. For the rounded vowels [u and o], on one hand, the trajectories indicating

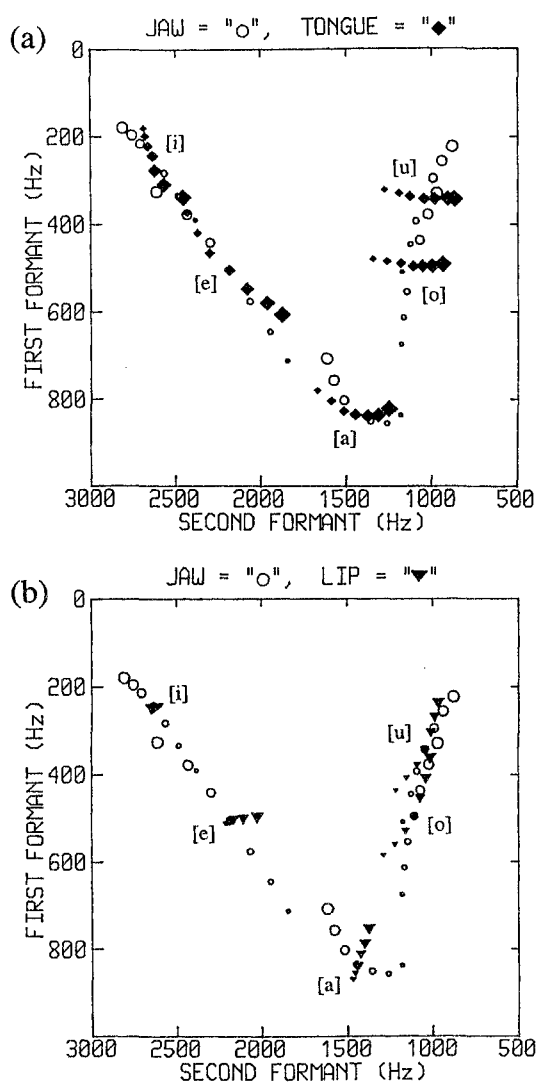


Figure 2. (a) Acoustic effects of a change in the jaw position marked by the circles and in the tongue-dorsal position by the diamonds, around the five cardinal vowels. (b) Same as (a), except acoustic effects of a variation in the lip aperture marked by the triangles are superimposed with those in the jaw positions.

the effects of the jaw parameter variation are almost vertical. Observe the dispersion of the open circles. The dispersion of the diamonds indicating the effects of the dorsal parameter variation, on the other hand, are almost horizontal. This means that for the rounded vowels the effect of the jaw appears on the F1 value alone, and those of the dorsal position primarily on F2.

For the unrounded vowels [i, e, and a], the F1-F2 dispersions of the two articulators tend to overlap each other. This means that a deviation in the position of one of the two articulators can be compensated by an adjustment of the remaining articulator to produce a "target" F1-F2 pattern of these unrounded vowels.

Interestingly the situation is exactly opposite for the jaw and the lip aperture (as the results of the lip rounding or spreading gesture). In Fig. 2b, the F1-F2 dispersions around the same five vowels as acoustic affects of the lip parameter, indicated by the closed triangles, are superimposed on these of the jaw. In contrast to the jaw-dorsal case, the overlap occurs for the rounded vowels [u and o], and the F1-F2 dispersions of the two gestures are roughly orthogonal for the unrounded vowel series, [i, e, and a]. A deviation of the jaw position therefore can be acoustically compensated for any vowel either by the adjustment of the dorsal position or by that of the lip rounding, depending on the two vowel categories.

It is noted that the F1 and F2 frequencies are quite stable against the variation in the lip aperture, in particular for the vowel [i], where the triangles are closely clustered together. This does not mean no effect of the rounding gesture on that vowel. The effects of the rounding can appear as a lowering of F3 frequency. The rounding puts the F3 closer to F2 and apart from F4, resulting in the sound change from [i] to [y].

6. COMPENSATORY ARTICULATION

The articulatory-acoustic relationships have indicated that the compensation (or one might prefer the term, inter-articulatory coordination) between the jaw and tongue-dorsal positions is acoustically effective for the unrounded vowels, and between the jaw and lip positions for the rounded vowels. One might raise the question whether or not speakers actually exploit the compensatory possibility during the production of speech. In order to provide an answer to this question, we have calculated the frame-by-frame variation of the articulatory parameters for the 10 sentences pronounced by the two speakers. The behavior of parameter trajectories during vowels is then investigated. Unfortunately, the number of the occurrences of the same vowel is small, ranging from only once to at most seven times. The data for only two vowels, [i and a], were used for the investigation. As a consequence of this, only the jaw-dorsal coordination is tested.

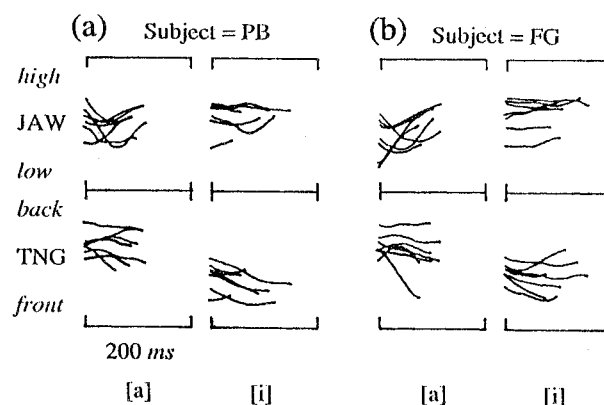


Figure 3. Superimposed trajectories of the jaw and tongue-dorsal parameters during the vowels [i and a] in different phonetic contexts, for the speaker PB at the left-hand and DF at the right-hand.

The calculated trajectories of the jaw and dorsal parameters concerning the two vowels are superimposed by aligning them at each vowel onset, as shown in Fig. 3. They exhibit a large variability for the same vowel, which are in different phonetic contexts. It is noted that the vertical space between the two abscissas for each parameter equals to six times the observed standard deviation, which corresponds roughly to the maximum range of the parameter variation. What's more striking however is the relatively close range of the two bundles of the trajectories corresponding to these two extreme vowels. As seen in the upper part of Fig. 3, the vertical ranges of the jaw parameter trajectories of these vowels overlap considerably each other in particular for the speaker PB. At the lower part of the figure, the dorsal position tends to be more fronted for [i] than [a], but the ranges of the variations still overlap to some extent.

Even though we do not expect an invariant target for individual articulators, it is still surprising to find such a poor contrast between high and low vowels. Indeed, the figures nicely fit with the prediction from the articulatory-acoustic relationships described above; such a magnitude of the variations and the poor contrast is possible, because the acoustic consequences can be compensated by the coordination between the jaw and the dorsal gesture. If this is the case, the two parameters should exhibit a linear or at least a monotonous relation for each vowel.

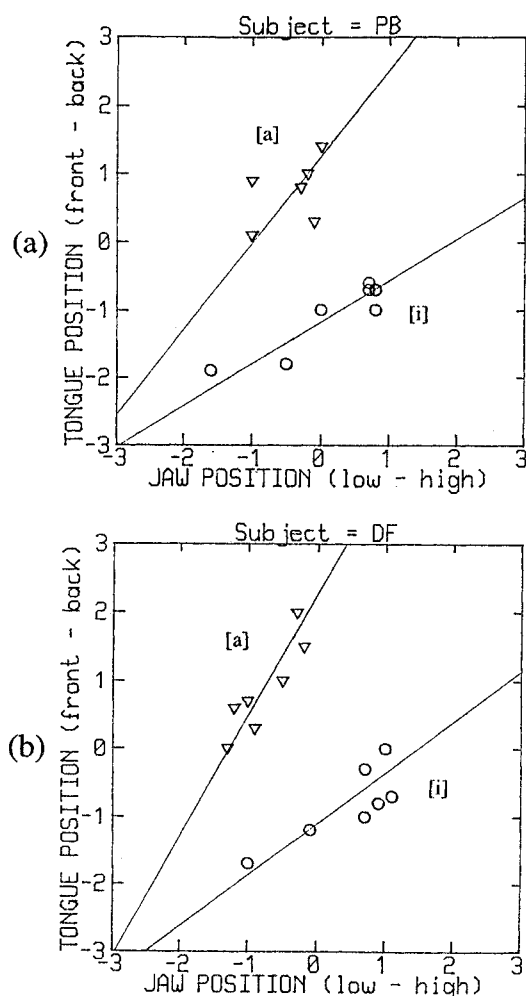


Figure 4. A linear relationship, indicated by the straight lines, between "target" positions of the jaw on abscissa and the tongue-dorsal on the ordinate, for the two unrounded vowels [i and a]; the speaker PB at (a) and DF at (b).

In order to verify this hypothesis, the tongue-dorsal positions are plotted as a function of the corresponding jaw positions, as shown in Fig. 4a for the speaker PB. The dorsal and jaw positions are sampled at minimum or maximum point of the contours within each vowel token. In the case where there is no such point, the parameter values are sampled at a center of the vowels. The jaw-dorsal plots of each vowel are indeed distributed along the corresponding straight lines drawn by visual inspection. The data concerning the second speaker DF are shown in Fig. 4b. The linear relation is more pronounced for this speaker, supporting the prediction.

7. CONCLUDING REMARKS

The speaker's strategy in the articulation of sentences can be quite complex. With the possibility of the compensatory articulation, it is not necessary to postulate fixed targets for the individual articulators, and thus there is a certain degree of freedom in the articulatory strategy, which can be exploited, for example, to minimize articulatory efforts. In other words, the articulatory strategy can have multiple solutions. It is then possible that every speaker has different solutions even under the constraint such as the minimal effort. If this is the case, every speakers might

articulate the same sentence differently. Consequently, the line of study described here will become excessively complex in trying to relate the articulations, *i.e.*, the coding process, with the phonetic universal, which is essentially speaker independent. A preliminary indication from the present data is in opposition to such a chaotic manner in articulation, however.

When calculated parameter movements for the same sentence pronounced by the two speakers are compared, we are rather surprised to find more similitude than difference between them. This means that although there are multitude of different ways to produce the same sentence, both speakers use a similar strategy to pronounce it, suggesting a relatively invariant rule issuing commands to the articulators. However we do not expect a unique set of such rules for a given language. Probably there are many different set of rules. Without such multiplicity, it is difficult to explain why there are so many different dialects and idiolects within the same language group.

We feel that a larger body of data is needed to investigate further the question of the articulatory strategy, or simply the question of how people speak. In our hope, the work presented here has demonstrated convincingly that the analysis of the conventional cineradiographic data using an articulatory model can provide a deeper insight into the speech production process, and that it can help us to understand better the inherent characteristics of speech, the variability.

ACKNOWLEDGMENTS

The author is grateful to Péla Simon, André Bothorel, Jean-Pierre Zerling, and François Wioland at l'Institut de Phonétique de Strasbourg for providing us with the digitized x-ray data base, as the result of a scientific collaboration between the institute and CNET.

REFERENCES

- Bothorel, A., Simon, P., Wioland, F., et Zerling, J. P. (1986). *Cinéradiographie des Voyelles et Consonnes du Français*, Travaux de l'Institut de Phonétique de Strasbourg.
- Gay, T., Lindblom, B., and Lubker J. (1981). "Production of bite-block vowels: Acoustic equivalence by selective compensation," *J. Acoust. soc. Am.*, 69, 3, 802-810.
- Lindblom, B. E. F. and Sundberg, J. E. F. (1971). "Acoustical consequences of lip, tongue, jaw, and larynx movement," *J. Acoust. Soc. Am.*, 4 (Part 2), 1166-1179.
- Lindblom, B., Lubker, J., and Gay, T. (1979). "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," *J. Phonet.* 7, 147-161.
- Maeda, S. (1978). "Une analyses statistique sur les positions de la tongue: Etude préliminaire sur les voyelles françaises," 9èmes JEP, GALF, 191-199.
- Maeda, S. (1979). "Une modèle articulaire de la tongue avec des composantes linéaires," 10èmes JEP, GALF, 152-164.
- Majid, R., Boë, L.J., and Perrier, P. (1986). "Fonctions de sensibilité, modèle articulaire et voyelles du français," 15èmes JEP, GALF, 59-63.
- Öhman, S. E. F. (1966). "Coarticulation in VCV utterances: Spectrographic measurements," *J. Acoust. Soc. Am.*, 39, 151-168.
- Overall, J. E. (1962). "Orthogonal factors and uncorrelated factor scores," *Psychological Reports*, 10, 651-662.
- Vaissière, J. (1987). "The use of allophonic variations of /a/ in automatic continuous speech recognition of French," Speech Communication Group Working Papers, Research Laboratory of Electronics, Massachusetts Institute of Technology, Volume V, 15-25.