MODELLING ARTICULATORY COMPENSATION FOR SYNTHESIS BY RULE

G.Boulianne[1,2], H. Cedergren[1] & D. Archambault[2]

[1]Université du Québec à Montréal, Canada
[2]INRS-Télécommunications, Québec, Canada

ABSTRACT

In this paper we propose a computer model of articulatory compensation such as needed in articulatory synthesis by rule. Targets are specified as acoustically important area-function features instead of formant frequencies. Articulatory space can then be searched for the articulatory position which, once translated into area-function features, will minimize distance to target. Search is constrained to physiologically possible positions; interarticulator dependencies and articulatory effort are taken into account. The model's behavior is shown to parallel that of real speakers in vowel production and bite-block experiments.

1. INTRODUCTION

In articulatory synthesis by rule, one is given a phonemic description and has to find the corresponding articulator positions of a computer model, so that its acoustic response can be computed. Realistic physiological/acoustical models can achieve phonemically equivalent productions using many different articulator configurations. The problem is then to find a strategy to choose among all the possible articulatory alternatives.

Human speakers are faced with the same problem; in fact they routinely exploit this freedom to closely approximate intended formant frequency targets during normal speech [8] or when an articulator is artificially constrained [5][6].

Previous attempts at computing articulatory positions from formant frequencies, however, have proved to be very costly in computation time and difficult because of the number of locally-only optimum solutions [1][4].

Although coding the target in terms of acoustically important area-function features has been proposed as an explanation for speaker behavior [5], it has never been used as a mean to derive articulatory positions. Computational complexity can be greatly reduced since there is no need for the time-consuming step of computing a vocal tract acoustic response. In the next sections we will show how area-function features can be used as the basis of a compensation model, and compare such a model with observed speaker behavior.

2. AREA-FUNCTION FEATURES

We first simplify the area-function to a concatenation of four cosinusoïdal elements. Acoustically critical dimensions are chosen as features: LT (total length of vocal-tract), OL (area at the lips), XC (distance between glottis and constriction), AC (area at the constriction), MG (maximum area between glottis and constriction) and ML (maximum area between constriction and lips) as illustrated in Figure 1. Similar sets of vocal-tract area-function features have been used by [1], [2] and [4].

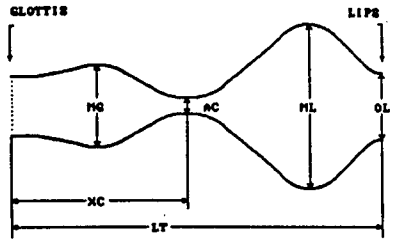

Figure 1. Area-function features

3. ARTICULATORY MODEL

In order to simulate inter-articulator dependencies and limit the search to physiologically realistic positions, we implemented the articulatory model of [7]. Only six major articulatory parameters were used: AM (jaw opening), AC and DC (tongue body position), AA (tongue tip angle), EL (vertical lips separation) and AL (lips protrusion). Most of the parameters are not absolute positions but measure articulator displacement relative to others (see [7]).

Given specific values for the parameters, the articulatory model generates a sagittal contour and, from it, area of sections along the length of the vocal-tract. Features are then derived from this area-function. The whole process is quite computation intensive, due to the articulatory model complexity. We developed a polynomial approximation to that process, such that area-function features are computed directly as a weighted sum of polynomial combinations of the articulatory parameters. This approximation has the benefit of smoothing out discontinuities in the articulatory-to-area-features relationship - that could prevent the search from reaching global optimum - as well as providing a tenfold reduction in computing time.

4. COMPENSATION STRATEGY

The model strategy is to seek for an optimal articulatory position, that is, one which translated into area-function features will be closest to the feature target. The choice of a distance measure that defines "closeness" has a profound influence on the final behavior of the model.

4.1. Distance measure

We selected a simple weighted euclidean distance: the sum of the squared differences between features, where each feature is first divided by its standard deviation over a training set.

By itself, however, this distance measure makes no difference between "easy" and "difficult" to reach positions, and could accept unreasonable positions as good solutions. Modifying the distance measure to take articulatory effort into account can be done by simply adding to it a sum of the squared differences between each articulator and its rest position, where each articulator is first divided by its rest position.

Note that both "distance to feature target" and "effort" components of the total distance are expressed as ratios. Using such dimensionless units avoids introduction of arbitrary weighing coefficients, that would otherwise be needed to adjust each component's contribution to the total distance.

4.2. Validation of effort component

An experiment was run to ascertain the effectiveness of the "effort" component in producing positions resembling those which speakers would choose. We compared solutions obtained using this distance measure with radiographic evidence from French speakers. Figure 2 plots jaw angles given by the model as a function of incisor distance measurements made on radiographic data from continuous speech [9]. To obtain the jaw angles the compensation model was fed with features computed from the area-function data of [3]. The correlation of

79% with radiographic data is quite impressive, given the fact that the area-function features contained no specific information about jaw angle: the agreement between simulated and real data comes solely from the use of the "effort" component in the distance measure. Correlations for other articulatory parameters that have been measured on radiographic data in [9] are 83% for vertical lips separation and 88% for lips protrusion. In these results, both the "distance" and "effort" components come into play, since area features OL and LT do contain some information about lips position.

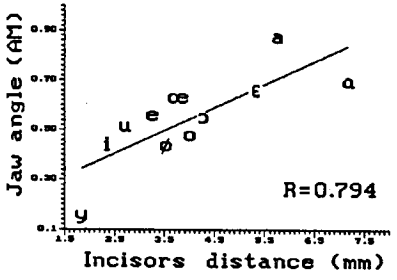

Figure 2. Vowel opening for computer and radiographic data

5. EXPERIMENTAL RESULTS

By constraining the jaw parameter to a single constant value during the search, we ran the computer equivalent of a bite-block experiment. Using the computer model and area-function data from [3], we obtained articulatory positions for the vowels /i,a,u/ in three cases:

1. *Normal case*: no articulator constraints.
2. *Compensated case*: imposing a specific "far from natural" jaw angle value while allowing other articulators free movement.
3. *Uncompensated case*: using articulators obtained in condition 1 and moving only the jaw to the imposed angle of condition 2. This shows the effect of the bite-block as if no compensation had taken place.

For /a,u/ the imposed jaw angle was AM=0.25 (equivalent to a 5.5 mm incisors distance), and for /i/ it was AM=0.80 (22 mm incisors distance). These conditions are similar to those used in [5] for human speakers.

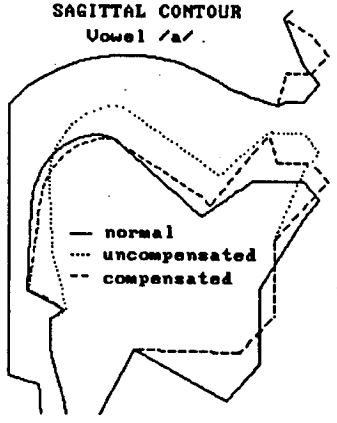

Figure 3. Articulators in bite-block experiment

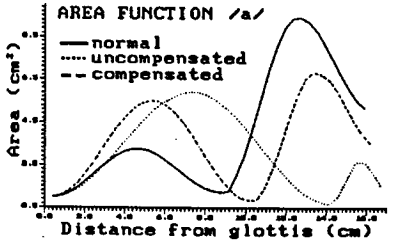

Figure 4. Area-function in bite-block experiment

For all three vowels area-function features in the uncompensated case are far from the normal case (average difference 66%), while values in the compensated case are close to the normal case (average difference 17%). Worst compensation occurred for vowel /a/. Figure 3 shows /a/ articulatory positions, and Figure 4 corresponding area-functions. Note that compensation is not perfect because the effort would be too high. Main observations of [5] on human speakers also applies to our data:

1. Compensation is attained by "supershaping of the tongue relative to its attachments to the jaw" [5], in our case affecting parameters that define tongue body position relative to the jaw. A lip parameter (jaw-relative vertical lip position) is also affected.

2. Area and position of main constriction are better preserved (average difference 17%) than front and back cavity areas (average difference 25%).

3. Area at the lips is better preserved in /u/ than /i/ or /a/ (differences of respectively 5%, 28% and 33%).

4. Acoustic response computed from the area-functions shows that the first two formant frequencies in the compensated case approximate well those of the normal case (average difference 11%), while formants in the uncompensated case would be far from normal (average difference 40%).

6. CONCLUSION

Our model simulates articulatory compensation as an optimality seeking process. Coding production targets as acoustically-important area-function features is efficient and reproduces speaker behavior in bite-block vowels experiments. Adding an effort component insures that articulatory displacement trade-offs that occur during continuous speech vowel production are also correctly simulated.

As it stands, this model is currently used for both consonant and vowel production, but has only been validated for vowels. Dynamic effects like coarticulation are not modelled, but can easily be included by adding distances to past and future positions.

7. REFERENCES

[1] ATAL, B.S., CHANG, J.J., MATHEWS, M.V., & J.W. TUKEY (1978) "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", *J.A.S.A.*, vol. 63, no 5, p. 1535-1555.

[2] COKER, C.H. (1976) "A Model of Articulatory Dynamics and Control", *Proc. IEEE*, vol. 64, no 4, pp. 452-460.

[3] FENG, G. (1987) "Etude articulatori-acoustique des voyelles nasales du français", *Bulletin de l'Institut Phonétique de Grenoble*, vol. 16, p. 1-102.

[4] FLANAGAN, J.L., ISHIZAKA, L., SHIPLEY, K.L. (1980), "Signal models for low bit-rate coding of speech", *J.A.S.A.*, vol. 68, no 3, pp. 780-791.

[5] GAY, T., B. LINDBLOM & J. LUBKER (1981) "Production of Bite-Block Vowels : Acoustic Equivalence by Selective Compensation", *J.A.S.A.*, vol. 69, no 3, pp. 802-810.

[6] LINDBLOM, B., J. LUBKER & R. MCALLISTER (1977) "Compensatory Articulation and the Modeling of Normal Speech Production Behavior", dans R. Carré, R. Descout & M. Wajskop, eds., *Modèles articulatoires et phonétiques*, Grenoble, pp. 147-161.

[7] MERMELSTEIN, P. (1973) "Articulatory Model for the Study of Speech Production", *J.A.S.A.*, vol. 53, no 4, pp. 1070-1082.

[8] PERKELL, J.S. (1989) "Testing Theories of Speech Production : Implications of some Detailed Analyses of Variable Articulatory Data", in Hardcastle & Marchal, eds., *Speech Production and Speech Modelling*, Kluwer, Dordrecht, 448 p.

[9] SIMARD, C. (1985) *Etude des séquences du type consonne constrictive plus voyelle en français, à l'aide de la radiocinématographie et de l'oscillographie*, CIRB, Québec, 403 p.