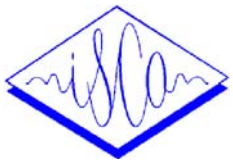

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A TRAJECTORY FORMATION MODEL OF ARTICULATORY MOVEMENTS USING A MULTIDIMENSIONAL PHONEMIC TASK

Tokihiko Kaburagi, Masaaki Honda, and Takeshi Okadome

NTT Communication Science Laboratories* and CREST/JST
*3-1, Morinosato-Wakamiya, Atsugi, Kanagawa, 243-0198 Japan

http://www.brl.ntt.co.jp/people/kaburagi/index.html

10.21437/Eurospeech.1999-34

ABSTRACT

This paper presents a model for representing context-dependent variation of articulator movements. Our model explains the contextual effect based on a multidimensional phonemic task and dynamic constraints of movements. The task determines the articulatory target so that invariant features of phoneme articulation are achieved. The dynamic constraints represent smoothly moving behavior of the articulators. Because the dimension of the task is smaller than that of the articulator variables, there are unconstrained degrees-of-freedom of the articulator variables. These redundant components are used to represent the contextual effect by smoothly interpolating the adjacent tasks. The phonemic invariant feature is defined as a linear transformation that minimizes a normalized articulatory variation. Simulation of articulatory movements is performed and the results are compared with actual movements.

1. INTRODUCTION

Context dependent coarticulatory phenomena, such as carry-over effects and anticipatory movements, are characteristic features of continuous speech utterances. These effects are the source of variability in the phoneme articulation. Based on the task-oriented trajectory formation approach, computational models of articulatory movements are proposed to study contextual variability [1,2,3]. The mechanism for generating contextual variability is explained in these models by means of the dimensionality of the task space and dynamic constraints regarding the smoothness of articulator movements. The dimension of the task space generally becomes smaller than that of the articulator variables because of the coordinated structure of the articulators and existence of priority among the articulators. Therefore, there are unconstrained degrees-of-freedom of the articulator variables and these redundant components are used to represent the contextual effect by smoothly interpolating the adjacent tasks.

This paper describes an alternative model directed at the problem of context-dependent variability. The articulatory target is represented in our model by statistically-defined phoneme-specific invariant features that minimize a normalized articulatory variation. Compared with vocal-tract tasks [1,3], they do not have clear physical meanings such as the size and location of the vocal-tract constriction. Instead, our phonemic task can flexibly represent the features of the phoneme articulation: movements making the vocal-tract constriction and relative movements among articulators. In addition, the dimension of the task space can be changed continuously to control the unconstrained degrees-of-freedom of the articulator variables. This property is suitable for explaining contextual articulatory movements.

In this paper, we first describe the mathematical representation of the phonemic task in section 2 and then, the formulation of the trajectory formation model is presented in section 3. Next, simulation is performed in section 4 to generate articulatory movements from input phoneme symbols and the results are compared with electro-magnetically measured actual movements.

2. TASK REPRESENTATION

The phonemic task represents gesture-based features of phoneme articulation that should be achieved by articulator movements. The closing movement of the lips for bilabial consonants or that of the tongue for dental and velar consonants are typical examples of such features. Another important feature is the relative movement (task-sharing) of the articulators: the jaw coordinates with the lower lip and tongue, and the body and tip of the tongue behave in a coordinated manner.

To represent these articulatory features, we define the task of a phoneme p by using a linear transformation $\mathbf{f}_p$ of articulator variables $\mathbf{x}$ as

$$0 = \mathbf{f}_p^t(\mathbf{x} - \bar{\mathbf{x}}_p) \qquad (1)$$

where $\bar{\mathbf{x}}_p$ is the mean vector. This linear transformation (phonemic invariant feature space) is determined so that it minimizes a normalized articulatory variance [4]

$$J(\mathbf{f}_p) = \frac{\mathbf{f}_p^t \Sigma_p \mathbf{f}_p}{\mathbf{f}_p^t \Sigma_T \mathbf{f}_p} \quad (2)$$

where Σ_p is the within-class covariance matrix of the articulatory variables and Σ_T is the total covariance matrix. Linear transformation $\mathbf{f}_p$ is obtained by solving the following generalized eigenvalue problem:

$$\Sigma_p F_p = \Sigma_T F_p \Lambda_p \quad (3)$$

where $F_p = (\mathbf{f}_{p1}, \mathbf{f}_{p2}, ..., \mathbf{f}_{pL})$ is the eigenvector matrix and $\Lambda_p = diag(\lambda_{p1}, \lambda_{p2}, ..., \lambda_{pL})$ represents the eigenvalue matrix ($\lambda_{p1} < \lambda_{p2} < ... < \lambda_{pL}$). L is the dimension of the articulator variables. The invariant feature space is finally determined as the subspace $\tilde{F}_p = (\mathbf{f}_{p1}, \mathbf{f}_{p2}, ..., \mathbf{f}_{pL_p})$ where L_p is the subspace dimension ($L_p \leq L$).

The invariant feature space described above represents characteristic features of the phoneme articulation that are less variant across phoneme context or other utterance conditions. Therefore, the phonemic task can be defined, as expressed in Eq. (1), by setting the projection of the articulator variables into the feature space as zero. In the feature space, movements for making a vocal tract constriction are represented by transforming the horizontal and vertical variables of an articulator into an axis that is perpendicular to the shape of the palate. Also, relative movements among articulators, such as coordination and interdependency reflecting the physiological structure, can be represented by transforming the variables of different articulators.

Each component of the feature space is ordered by the eigenvalue which is equivalent to the normalized variance J and thus represents a degree of articulatory consistency. Therefore, the dimension of the phonemic task (L_p) can be set according to this consistency to control the unconstrained degrees-of-freedom of the articulator variables. When the task dimension is small, there are many degrees-of-freedom of the articulatory variables that can be used to explain the contextual variability. To the contrary, the position of every articulator is fixed to the mean vector when the task dimension is the same as that of the articulator variables.

The scatter plot in the left part of Fig. 1 shows the position of the jaw (J), upper lip (UL), lower lip (LL), tongue (T1, T2, T3, and T4), and velum (V) in the articulation of five phonemes. In the right part, the first component of the invariant feature space is displayed by plus marks, calculated by multiplying several gain factors to the eigenvector and adding it the mean vector. From the direction of the feature space, it is clear that the feature space emphasizes an articulator behavior making the vocal tract constriction for

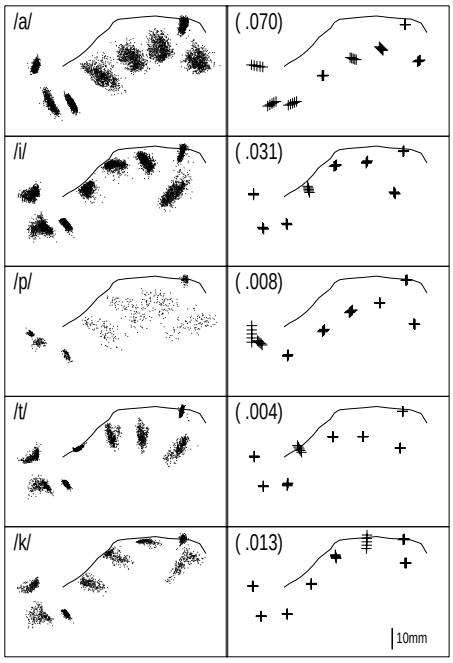

Figure 1: Scatter plot of the articulator position and the first component of the phoneme invariant feature space.

the vowel /i/ and plosive consonants. The eigenvalue is smaller for consonants than vowels indicating that the articulator position of the consonants is less variant. The articulatory data set was obtained using an electro-magnetic articulograph system (Carstens AG100) [5,6] by measuring the position of the articulators during utterances of 354 sentences.

3. TRAJECTORY FORMATION MODEL

Figure 1 illustrates the outline of the trajectory formation model. For input phoneme symbols, phonemic task as well as the articulatory timing is specified. The trajectory of articulatory movements is determined so that it satisfies given phonemic tasks and two types of dynamic constraints.

3.1. Phonemic Task Sequence

For each input phoneme symbol, the phonemic task representing the characteristic articulatory gesture is specified using the invariant feature space. The task representation given in Eq. (1) constructs linear simultaneous equations with respect to the unknown articulatory variables $\mathbf{x}$. The order of these equations is the same as the dimension of the feature space L_p. Therefore, the phonemic task sequence can be written as

$$\mathbf{z}_k = E_k \mathbf{x}(n_k) \quad (1 \leq k \leq K) \quad (4)$$

where $\mathbf{z}_k$ represents the target vector and E_k is the transformation matrix. n_k represents the time instant

at which the phoneme should be articulated and k is an index of the task sequence. $n_1(=1)$ and $n_K(=N)$ correspond respectively to the initial and terminal ends of articulatory movement. In our model, the articulatory timing n_k should be specified.

3.2. Dynamic Constraints

Because the task dimension is smaller than that of the articulator variables, the phonemic task only constrains the partial degrees-of-freedom of the articulator variables. Thus the inverse mapping that determines $\mathbf{x}$ from $\mathbf{z}$ becomes one-to-many (ill-posed). On the other hand, the phonemic task sequence is temporally organized so that each task is achieved at a particular point in time. Articulators can then take an arbitrary position between adjacent motor tasks.

To resolve these redundancies and represent the smoothly moving behavior of the articulators, two types of the dynamic constraint are introduced. First, the trajectory of each articulatory variable is represented as the output of a linear second order dynamic system:

$$x(n) - 2\tau x(n-1) + \tau^2 x(n-2) = (1-\tau)^2 y(n) \quad (5)$$

where y represents the input force to the system and τ is a model parameter. Secondly, an objective function representing the smoothness of articulatory movements is defined as [7]

$$C = \sum_{n=1}^{N-1} (C_x(n) + C_y(n)) + C_x(N) \quad (6)$$

where

$$C_x(n) = \| \mathbf{x}(n) - \mathbf{x}(n-1) \|^2_{(GW_xG^t)} \quad (7)$$

represents the energy criterion with respect to the velocity of movements and

$$C_y(n) = \| \mathbf{y}(n) - \mathbf{y}(n-1) \|^2_{(GW_yG^t)} \quad (8)$$

represents the criterion with respect to the change of system inputs. The total energy C is calculated by summing the instantaneous power related to the change of system inputs and outputs over the entire movement.

Weighting matrices of the objective function are set as $W_x = diag\{w_1, w_2, ..., w_L\}$ and $W_y = diag\{d_1w_1, d_2w_2, ..., d_Lw_L\}$. The parameter w_l determines the cost weight between each articulatory variable and can be used to control the relative amplitude between articulators. The parameter d_l determines the relative cost between C_x and C_y and is used to adjust the overall shape of the trajectory together with the parameter τ. G is an orthogonal matrix constructed from the eigenvectors of the total covariance matrix Σ_T. The total cost is weighted to each principal component of the articulator variables according to its variation by means of this orthogonal transformation.

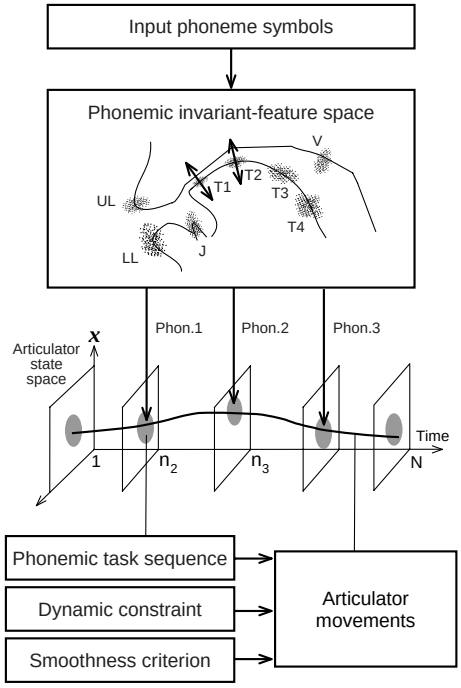

Figure 2: Outline of the trajectory formation model.

3.3. Determination of Articulatory Movement

The trajectory of articulatory movements is determined so that it satisfies the phonemic tasks sequence (Eq. (4)) and minimizes the objective function C under the constraint of the dynamic system representation (Eq. (5)). This framework produces an optimal control problem with linear dynamics and quadratic criteria. This problem can be solved explicitly by dynamic programming [3] when the values of the model parameters are fixed.

4. EXPERIMENT

4.1. Contextual Variability

Figure 3 shows the simulated trajectory of VCV sequences. The tongue shape of each consonant is shown as polygonal lines by concatenating the positions of four tongue points. The shape of the tongue is influenced by the preceding vowel while the movements for making the vocal tract closure are maintained. This simulation demonstrates the ability of the model to represent the contextual variability.

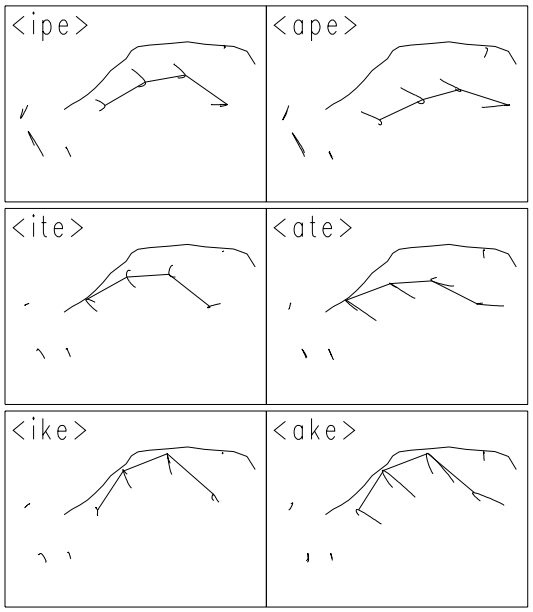

Figure 3: Simulated movements for VCV sequences.

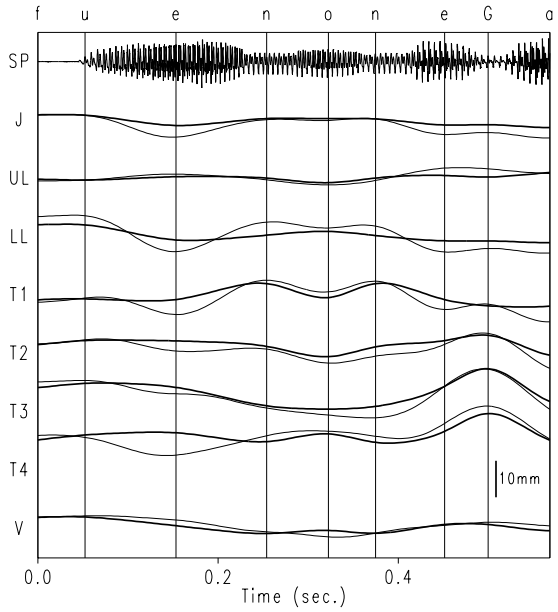

Figure 4: Comparison of measured and simulated movements.

4.2. Prediction Accuracy

To examine the prediction accuracy of the model, articulatory movements were generated and compared with measured movements. Articulatory data set was first separated into closed and open sets and the phonemic feature space was calculated using the closed set. Also, task dimensions and model parameters were determined using a search technique so that the mean error between the simulated and measured movements was minimized. Then simulation was performed for 24 sentences in the open data set while swapping the closed and open set. The mean simulation error for 48 sentences was 1.58 mm. This error was larger than that of the triphone task model [8] (1.21 mm) and was almost the same as that of the diphone model (1.48 mm). When the task was specified by the mean vector of each phoneme, the error was 1.76 mm.

Figure 4 compares measured and simulated movements. The horizontal axis is the time and traces show the speech waveform and vertical movements of the articulators. Thin and thick lines correspond respectively to measured and simulated articulator movements. Vertical lines indicate the articulatory timing of each phonemic task.

5. CONCLUSION

A model of articulator movements was presented to explain the context-dependent articulatory variability using context-independent phonemic tasks. The task was defined by an invariant feature subspace representing consistent articulatory gestures. The experimental results showed that our model is useful for representing the articulatory variability in continuous speech utterances.

REFERENCES

[1] Saltzman, E. and Munhall, K.G. (1989). "A dynamical approach to gestural pattering in speech production," Ecological Psychology **1**, 333-382.

[2] Bailly, G., Laboissiere, R., and Schwartz, J. L. (1991). "Formant trajectories as audible gestures: An alternative for speech synthesis," J. Phon. **19**, 9-23.

[3] Kaburagi, T. and Honda, M. (1996) "A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes," J. Acoust. Soc. Am. **99**, 3154-3170.

[4] Honda, M. and Kaburagi, T. (1996) "Statistical analysis of a phonemic target in articulatory movements," ASA and ASJ Third Joint Meeting 1pSC4.

[5] Kaburagi, T. and Honda, M. (1994) "Determination of sagittal tongue shape from the positions of points on the tongue surface," J. Acoust. Soc. Am. **96**, 1356-1366.

[6] Kaburagi, T. and Honda, M. (1997) "Calibration methods of voltage-to-distance function for an electro-magnetic articulometer (EMA) system," J. Acoust. Soc. Am. **101**, 2391-2394.

[7] Okadome, T. and Honda, M. (1992). "Trajectory formation in sequential arm movements," Proc. IEEE Conf. SMC, 471-478.

[8] Okadome, T., Kaburagi, T., and Honda, M. (1998) "Trajectory formation of articulatory movements for a given sequence of phonemes," Proc. ICSLP98.