Statistical identification of critical, dependent and redundant articulators

Veena D. Singampalli and Philip J.B. Jackson

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, U.K.
v.singampalli@surrey.ac.uk and p.jackson@surrey.ac.uk

Abstract

A compact, data-driven statistical model for identifying roles played by articulators in production of English phones using 1D and 2D articulatory data is presented. Articulators critical in production of each phone were identified and were used to predict the pdfs of dependent articulators based on the strength of articulatory correlations. The performance of the model is evaluated on MOCHA database using proposed and exhaustive search techniques and the results of synthesised trajectories presented.

Index Terms: coarticulation, speech production, articulatory modeling, critical articulator

1. Introduction

Coarticulation is one of the main problems faced by researchers in speech technologies. Many researchers tried to explain the invariance associated with phonetic segments in acoustic and articulatory domains using features, targets or goals. They viewed coarticulation as a spread of features [8, 15, 6] or coproduction of articulatory gestures [20, 4, 14, 12]. In feature based models, a set of binary features [5, 7] were specified for each phone based on phonological rules. Formant based targets for some vowels and consonants were presented in [13, 16]. Phonetic invariance in articulatory space was also explained using critical articulator concept based on phonetic rules using degree of articulatory constraint [18] and coarticulatory resistance [3]. Later on, powerfull statistical models which make efficient use of the available data took over the rule based ones, and models such as segmental HMMs [19] and trajectory HMMs [21] have been proposed in recent years to capture the dynamics of speech. The variation due to context was modelled in a statistical way using articulatory curvatures of preceeding and following contexts [2]. Our attempt to identify the phonetic invariance in the articulatory domain in an entirely statistically driven way is presented in this paper. Effective models of the constraints of the human articulatory system in speech production have the potential to ensure naturalness in speech synthesis and to improve speech recognition in noisy conditions.

Researchers have used different kinds of articulatory data, such as MRI, X-ray cine and microbeam, electro-palatograph (EPG), electro-glottograph (EGG) and electro-magnetic articulograph (EMA) [17, 22, 23]. The MOCHA articulatory database [23] contains EPG, EGG, EMA and audio recordings of male and female speakers uttering 460 TIMIT sentences each. With coils on the bridge of the nose and upper front incisor providing reference, there are 14 EMA channels of x and y movements for 7 articulators: upper lip UL, lower lip LL, lower incisor LI, tongue tip TT, tongue blade TB, tongue dorsum TD, and velum V. Fig. 1 gives a midsaggital display of articulators, showing the outline of normal distributions fitted to data for two phones, [s] and [g].

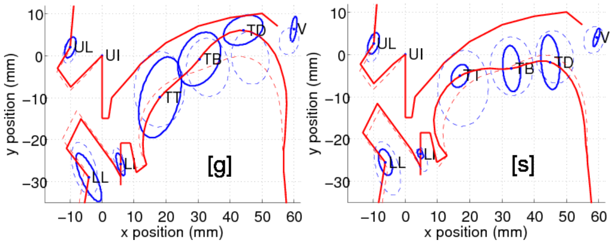


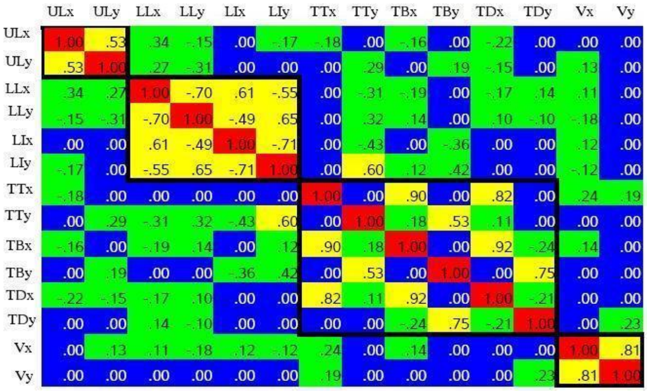
Figure 1: *Midsaggital display of different articulatory points and distributions of global (dotted) and phone specific distributions (solid) of [g] and [s] generated using male speaker (msak)*

During an utterance, the articulators continually change their role. If an articulator has to attain a specific position or a gestural movement is essential for production of a speech sound, such articulator is considered to be **critical** for that sound, e.g. position of tongue tip in production of alveolar fricative [s] in fig. 1. An articulator is said to be **dependent** on one or more critical articulators if it follows the critical articulators and its position is influenced by them. A **redundant** articulator is not constrained during the production of a speech sound and its position does not effect the production of sound critically. As an example, movement of tongue dorsum towards the velum during the utterance of a velar sound [g] is considered to be critical, tongue blade and tip are dependent, lips and jaw are redundant as shown in fig.1.

In our approach to model the movements of articulators during speech, we use the information of critical, dependent and redundant articulators and incorporate three different kinds of correlations associated with them (1) correlated movements of the articulators in space, (2) correlations amongst the articulators and (3) correlations over time. The first kind of correlation is most apparent with the jaw and lower lip that move vertically much more than in the anterior-posterior direction. The second kind captures the biomechanical relationship between articulators, for example, amongst different points on the tongue. The third kind of correlation mentioned here is associated with the smooth movements of the articulators over time as in [21] [2] [10]. The present model incorporates the first two correlations in its structure in an entirely data driven, statistical and compact way and uses some of the existing methods to generate smooth trajectories. The following sections in this paper present our model description, evaluation of the models' performance, conclusions and future work.

2. Modeling articulator roles

We propose an algorithm for identifying the set critical articulators for each phone statistically, using their locations to predict those of dependent articulators based on correlation. High



	ULx	ULy	LLx	LLy	LIx	LIy	TTx	TTy	TBx	TBy	TDx	TDy	Vx	Vy
ULx	1.00	.53	.34	-.15	.00	-.17	-.18	.00	-.16	.00	-.22	.00	.00	.00
ULy	.53	1.00	.27	-.31	.00	.00	.00	.29	.00	.19	-.15	.00	.13	.00
LLx	.34	.27	1.00	-.70	.61	-.55	.00	-.31	-.19	.00	-.17	.14	.11	.00
LLy	-.15	-.31	-.70	1.00	-.49	.65	.00	.32	.14	.00	.10	-.10	-.18	.00
LIx	.00	.00	.61	-.49	1.00	-.71	.00	-.43	.00	-.36	.00	.00	.12	.00
LIy	-.17	.00	-.55	.65	-.71	1.00	.00	.60	.12	.42	.00	.00	-.12	.00
TTx	-.18	.00	.00	.00	.00	.00	1.00	.00	.90	.00	.82	.00	.24	.19
TTy	.00	.29	-.31	.32	-.43	.60	.00	1.00	.18	.53	.11	.00	.00	.00
TBx	-.16	.00	-.19	.14	.00	.12	.90	.18	1.00	.00	.92	-.24	.14	.00
TBy	.00	.19	.00	.00	-.36	.42	.00	.53	.00	1.00	.00	.75	.00	.00
TDx	-.22	-.15	-.17	.10	.00	.00	.82	.11	.92	.00	1.00	-.21	.00	.00
TDy	.00	.00	.14	-.10	.00	.00	.00	.00	-.24	.75	-.21	1.00	.00	.23
Vx	.00	.13	.11	-.18	.12	-.12	.24	.00	.14	.00	.00	.00	1.00	.81
Vy	.00	.00	.00	.00	.00	.00	.19	.00	.00	.00	.00	.23	.81	1.00

Table 1: *Univariate correlation matrix R^* of strong and statistically-significant correlations.*

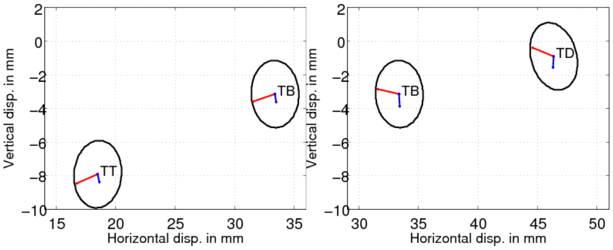


Figure 2: *Directions of first and second significant canonical correlations between* TT, TB *(left) and* TB, TD *(right), grand covariance ellipses are plotted in black.*

correlations between articulators induce strong effects from the critical articulators' configuration on the others. The algorithm seeks to explain the position distributions of all the articulators from data in terms of a compact set of critical positions, representing the essential articulation for a given phone. From various distance measures considered for testing the quality of the distribution match (student's t, Hotelling's T^2, Fischer's linear discriminant), Kullback-Leibler (KL) divergence was chosen [11], which measures the distance between pdfs in terms of their mutual information. Before we can identify the critical articulators, and examine their effect on the other articulators, we need first to find the significant correlations between them.

2.1. Correlation analysis between articulators and in space

EMA data for one male (msak) and female (fsew) speaker were smoothed using a Hann window and sampled at 10 ms frame rate. The model was implemented for 1D data, where x and y coordinates were assumed independent ($n = 14$), and for 2D data, where correlations in an articulator's x and y movement were included ($n = 7$). Global distributions over all utterances by each speaker modeled by univariate and bivariate normal pdfs respectively, based on the grand statistics: mean M_i and variance Σ_{ij} (1D); $\boldsymbol{M}_i$ and $\boldsymbol{\Sigma}_{ij}$ (2D).

Univariate model: Global 1D correlations were computed between the 14-channel articulatory data, $R = \{r_{ij}\}\ \forall i, j \in 1..n$. The Pearson significance test was applied at $\alpha = 0.05$, and statistically insignificant correlations were set to zero, along with very weak ones $|r_{ij}| < 0.1$. Table 1 shows the remaining univariate statistically significant correlations R^*, which are related to the covariance: $\Sigma_{ij} = \Sigma_{ii}^{1/2}\, r_{ij}\, \Sigma_{jj}^{1/2}$.

From table 1, we can see that tongue tip, blade and dorsum are highly correlated in the x direction (TT$_\text{x}$, TB$_\text{x}$, TD$_\text{x}$); important but less so in the y direction. The low correlations between x and y directions of TT, TB and TD highlight the tongue's independent movement in two directions. Strong correlations exist between lower lip and lower incisor (LL and LI). Upper lip (UL) is correlated with LL but there is little correlation between UL and the jaw LI, as expected; some correlation between TT and LI exists in the y direction. Velum, strongly correlated with itself in 2D, has almost no correlation with other articulators, so it does not suffer from inter-articulator gestural conflict.

Bivariate model: 2D correlations, taking x and y data together, were computed by canonical correlation (CC) analysis [9]. This technique finds the combined directions in which two articulators have maximum correlation and allows for the number of degrees of statistically-significant correlation to be tested:

$$\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_{ii}^{1/2}\, \boldsymbol{U}_i\, \boldsymbol{r}_{ij}\, \boldsymbol{U}_j'\, \boldsymbol{\Sigma}_{jj}^{1/2}, \tag{1}$$

where $\boldsymbol{U}_i$ and $\boldsymbol{U}_j$ contain the eigenvectors for each articulator, the 2×2 matrix $\boldsymbol{r}_{ij} = \operatorname{diag}\left(\rho_{ij}^1, \rho_{ij}^2\right)$ contains the CCs ρ_{ij} and $'$ denotes transpose. CCs were computed between every articulator pair i, j and significance tested at $\alpha = 0.05$. As before, weak and insignificant correlations were set to zero. Most articulatory pairs (56 %) had two significant CCs; very weak correlation was found between TD and LI; other articulator pairs had one significant correlation.

Figure 2 shows CCs between TT-TB, and TB-TD. The eigenvector directions emphasise the strong front/back correlation of motion tangential to the tongue surface ($\rho^1_{\text{TT,TB}}$=0.92; $\rho^1_{\text{TB,TD}}$=0.93), compared to raising/lowering ($\rho^2_{\text{TT,TB}}$=0.75; $\rho^2_{\text{TB,TD}}$=0.50). The bivariate correlations, expressed in the 14 articulator dimensions, were similar in absolute value to the univariate ones. However, the orthogonal representation of correlation compressed the total number of related degrees of freedom. The next section presents an iterative algorithm to identify critical articulators automatic for any given phone ϕ.

2.2. Identification and influence of critical articulators

1. Initialisation. Data samples at the phone midpoints provide reference statistics of normal articulatory distributions: mean μ_i^ϕ and variance σ_i^ϕ (1D); $\boldsymbol{\mu}_i^\phi$ and $\boldsymbol{\sigma}_i^\phi$ (2D).

The model pdfs are initialised with global statistics: $m_i^\phi = M_i$ and $s_i^\phi = \Sigma_{ii}$ (1D), $\boldsymbol{m}_i^\phi = \boldsymbol{M}_i$ and $\boldsymbol{s}_i^\phi = \boldsymbol{\Sigma}_i$ (2D).

2. Critical identification (C-step).

Identify critical articulator:

At each level, $k = 1..n$, calculate (univariate or bivariate) KL divergence between model and phone distributions $J_k^\phi(i)$ for each articulator i, incorporating the standard error by multiplying the variance of N data points by $(N+1)/N$. The next critical articulator $j = \operatorname{argmax}_i\{J_k^\phi(i)\}$ is selected iff $J_k^\phi(j) > \theta_\text{crit}$.

Set critical articulator distribution: Model statistics are set to phone statistics as 1D or 2D respectively:

$$m_j^\phi \mapsfrom \mu_j^\phi, s_j^\phi \mapsfrom \sigma_j^\phi; \qquad \boldsymbol{m}_j^\phi \mapsfrom \boldsymbol{\mu}_j^\phi, \boldsymbol{s}_j^\phi \mapsfrom \boldsymbol{\sigma}_j^\phi. \tag{2}$$

3. Dependent update (D-step).

Gather dependency statistics: The covariations are collated from the grand statistics, Σ_{ii}, Σ_{ij} and Σ_{ii}, for all dependent (iff $J_k^\phi(i) > \theta_\text{dep}$.) and critical articulators i and j (1D); $\boldsymbol{\Sigma}_{ii}$, $\boldsymbol{\Sigma}_{ij}$ and $\boldsymbol{\Sigma}_{jj}$ (2D).

Update dependent articulator pdfs: Combining these statistics

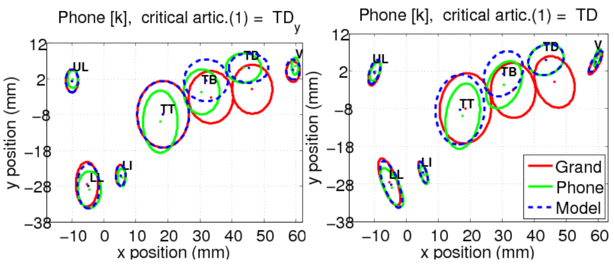


Figure 3: *Distributions of 1D (left) and 2D (right) models after identification of first critical articulator for phone [k],* $\theta_{\rm crit} = 0.5$ *and* $\theta_{\rm dep} = 0.1$.

with the phone ones for any critical articulator j, the conditional pdfs get re-estimated, according to [1], which for 2D case gives:

$$\begin{aligned} \boldsymbol{m}_i^{\phi} &\hookleftarrow \boldsymbol{M}_i + \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{jj}^{-1}\left(\boldsymbol{\mu}_j^{\phi} - \boldsymbol{M}_j\right) \\ \boldsymbol{s}_i^{\phi} &\hookleftarrow \boldsymbol{\Sigma}_{ii} + \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{jj}^{-1}\left(\boldsymbol{\sigma}_j^{\phi} - \boldsymbol{\Sigma}_{jj}\right)\boldsymbol{\Sigma}_{jj}^{-1}\boldsymbol{\Sigma}_{ij}'. \end{aligned} \quad (3)$$

The algorithm stops in C-step when $J_k^{\phi}(i) < \theta_{\rm crit} \forall i \in 1..n$.

3. Evaluation

3.1. Results on MOCHA

The results obtained after implementation of critical articulator detection algorithm on male and female speaker data from MOCHA database are presented in this section. The 1D and 2D model distributions are shown for phone [k] in fig.3, after the first critical articulator was identified at level k=1 in C-step and models updated in D-step. In the 1D case, the distribution of the first critical articulator, $\rm TD_y$, strongly influenced the distribution of $\rm TB_y$ as the correlation between them was strong (table 1). Since the correlations between $\rm LL_x$, $\rm LL_y$, $\rm V_y$ and $\rm TD_y$ were small in value, the distributions of those articulators were weakly affected. In the 2D case, the distribution of first critical articulator, TD, had a similar effect on configurations of TB and other articulators. The algorithm stopped after choosing 6 critical articulators ($\rm TD_y$, $\rm V_x$, $\rm TB_y$, $\rm UL_y$, $\rm LL_y$, $\rm TT_x$) in the 1D case and 3 (TD, V, TB) in the 2D case respectively.

On average, 4 articulators were chosen as critical per phone for 1D and 2D models when $\theta_{\rm crit} = 0.5$ and $\theta_{\rm dep} = 0.1$. As a measure of fit of the model distributions to the phone distributions, at each level k, the average of $J_k^{\phi}(i)$ was computed across all phones and articulators for 1D and 2D models. The average KL divergence dropped by 90% of its initial value in the 1D case and 93% of its initial value in the 2D case after all critical articulators were identified for each phone which indicates a significant improvement in the model convergence. The set of critical articulators chosen by our algorithm for each phone was compared with the articulatory features given in [5]. The results of our algorithm are in broad agreement with the active articulators and the corresponding target regions that were specified for each phone. For example, our algorithm chose upper and lower lips, velum as critical articulators for a bilabial nasal stop [m]. For the high-front vowel [i:], tongue tip, dorsum and lower lip were chosen as critical.

This critical articulation detection model is compact as the model distributions can be effectively represented using the identity and the statistics of critical articulators (for each phone), grand distributions and grand articulatory correlations.

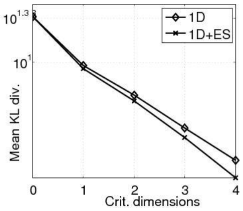


Figure 4: *Avg. KL divergence across all phones between 1D phone distributions and 1D models, estimated using proposed 1D algorithm (◇) and ES procedure (×).*

The number of parameters required for representing the model distributions using our algorithm was 40% less in the 1D case and 43% less in the 2D case when compared with the parameters in the conventional models where statistics of all articulators for every phone are stored.

3.2. Exhaustive search-1D models

An **exhaustive search** (ES) algorithm was implemented on 1D data to check if the set of critical articulators identified by the proposed algorithm gave the best model convergence and if their order affected the final models. A search for best critical articulator combination was performed at each level considering all possible combinations, independent of the previous critical articulators' information. The dependent articulator distributions were updated conditioned on each critical articulator combination (D-step). For every phone, the combination of critical articulators that gave minimum KL divergence between 1D model and phone distributions was chosen as critical at each level. The ES procedure was implemented upto 4 levels on male speaker data. We found that the order of critical articulators made no difference to the model convergence in all cases upto level k=4 except for 10% of the cases (a mere 1% improvement was given by ES procedure) when similar set of critical articulators were identified by both procedures. When two different sets of critical articulators were chosen, the ES procedure gave a 13% improvement over the proposed algorithm. The average KL divergence values between 1D model distributions (estimated using proposed and ES procedures) and phone distributions computed across all phones at each level are depicted in fig.4. The time taken to run ES upto level 4 was 3×10^5 seconds whereas the proposed algorithm took 100 seconds for the same. On average across all levels, the critical articulators identified using ES procedure reduced the divergence by 12% over those identified using the proposed algorithm at the expense of significant increase in the computational effort.

3.3. Trajectory generation

The performance of the models was evaluated by generating synthetic trajectories using the information of articulatory roles given by the proposed algorithm. The positions of critical and dependent articulators at every phone midpoint were set to the mean of the corresponding model distributions and successive target positions were linearly interpolated. No target positions were specified for redundant articulators since they are not constrained to achieve any target positions. The trajectories thus generated by modeling redundancies were compared with the original articulatory trajectories and the trajectories generated by linear interpolation between static targets where positions of the redundant articulators were also set to their model means.

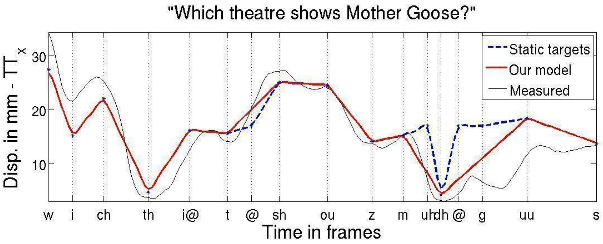


Figure 5: *Original and synthetic trajectories generated using static positions and our model for one sentence uttered by msak.*

All synthetic trajectories were filtered using a zero phase order 10 lowpass filter at 20Hz sampling frequency. Figure 5 depicts the movement of TT_x for one sentence uttered by the male speaker. For phones [uh], [@] and [g], TT_x was found to be redundant and therefore no target position was specified for TT_x at the midpoint of those phones. The trajectory generated by modeling redundancies was a closer fit to the actual trajectory than the one generated using static target positions. Normalised RMS errors and correlations between synthetic and original trajectories were used to evaluate the performance of the models. The normalised RMS error was reduced by 2.2% (male), 2.3% (female) and correlation was improved by 2.6% (male), 3.0% (female) when redundancy modeling was used over interpolation between static targets. Redundant articulatory positions were observed only in 17% of all the male speaker and 22% of all the female speaker data. Hence, the improvement was a small percentage when compared with modeling contextual effects of articulator acceleration on entire trajectories, as in [2], which gave 12.3% improvement on male and 8.7% on female data.

4. Conclusions and future work

We proposed a statistical algorithm for identifying the roles played by articulators in production of speech sounds using 1D and 2D articulatory data and evaluated the performance of the models on MOCHA database. The fit of the 1D and the 2D models to the respective phone distributions was presented. The list of critical articulators generated for each phone were found to be in broad agreement with the phonetic features. We found that the order of the critical articulators made no significant difference to the model convergence and the critical articulators obtained using an exhaustive search procedure improved the convergence by a small amount but with a considerable increase in the computational effort. Preliminary evaluation of trajectory generation using the information of articulatory roles gave positive results. Future work will focus on evaluating the performance of the models on other articulatory databases, using different trajectory generation schemes, finding ways to exploit redundant degrees of freedom within dependent articulators, and using the compact representation to improve articulatory speech recognition.

5. Acknowledgement

DANSA project is funded by EPSRC (GR/S85511/01) in UK, see http://www.ee.surrey.ac.uk/Personal/P.Jackson/Dansa/.

6. References

[1] T.W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, New York, 2 edition, 1984.

[2] S.C. Blackburn and S. Young. A self-learning predictive model of articulator movements during speech production. *JASA*, 103(3):1659–70, March 2000.

[3] R. A. W. Bladon and A. Al-Bamerni. Coarticulation resistance in English /l/. *J. Phon.*, 4:135–50, 1976.

[4] C.P. Browman and L. Goldstein. Towards an articulatory phonology. *Phonology*, 3:219–52, 1986.

[5] N. Chomsky and M. Halle. *The sound pattern of English*. Harper & Row, New York, 1968.

[6] R. Daniloff and R. Hammarberg. On defining coarticulation. *J. Phon.*, 1:239–248, 1973.

[7] G. Fant. Distinctive features and phonetic dimensions. *Applications of Linguistics, Cambridge, UK)*, 1969.

[8] W. L. Henke. *Dynamic articulatory model of speech production using computer simulation*. PhD thesis, MIT, Cambridge, MA, 1965.

[9] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*. Prentice Hall, New Jersey, 4 edition, 1998.

[10] P. A. Keating. The window model of coarticulation: articulatory evidence. *UCLA Working papers in Phonetics*, 69:3–29, 1988.

[11] S. Kullback. *Information theory and statistics*. Dover Pub., New York, 1 edition, 1968.

[12] A.M. Liberman. The grammars of speech and language. *Cog. Psych.*, 1:301–23, 1970.

[13] B. Lindblom. Spectrographic study of vowel reduction. *JASA*, 35:1773–81, 1963.

[14] P.F. MacNeilage. Motor control of serial ordering of speech. *Psychol. Rev.*, 77:182–196, 1970.

[15] K. Moll and R. Daniloff. Investigation of the timing of velar movements during speech. *JASA*, 50(2):678–84, 1971.

[16] S.E.G. Öhman. Coarticulation in VCV utterances: Spectrographic measurements. *JASA*, 39(1):151–68, 1966.

[17] V. Parthasarathy, J.L. Prince, M. Stone, E.Z. Murano, and M. NessAiver. Measuring tongue motion from tagged cine-mri using harmonic phase (harp) processing. *JASA*, 121(1):491–504, 2007.

[18] D. Recasens, D.M. Pallarés, and J. Fontdevilla. A model of lingual coarticulation based on articulatory constraints. *JASA*, 102:544–561, 1997.

[19] M.J. Russell and P.J.B. Jackson. Models of speech dynamics in a segmental-HMM recogniser using intermediate linear representations. *Proc. ICSLP,* Denver, CO, 2002.

[20] E.L. Saltzman and K. Munhall. A dynamic approach to gestural patterning in speech production. *Ecology Psychology*, 1:333–82, 1989.

[21] K. Tokuda, H. Zen, and T. Kitamura. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic vector feature sequences. *Comp. Speech & Lang.*, 21(1):153–73, 2007.

[22] J.R. Westbury, G. Turner, and J. Dembowski. *X-ray microbeam speech production database user's handbook.* University of Wisconsin, Madison, WI, 1994.

[23] A.A. Wrench. A new resource for production modelling in speech technology. *Proc. Inst. of Acoust.,* Stratford-upon-Avon, UK, 2001.