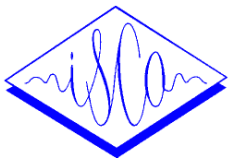

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

# PHONEME EXTRACTION USING VIA POINT ESTIMATION OF REAL SPEECH

*E. Vatikiotis-Bateson[1], M. Tiede[1], Y. Wada[2], V. Gracco[3], and M. Kawato[1]*

[1]ATR Human Information Processing Research Laboratories, Kyoto, Japan
[2]Kawasaki Steel Company, Chiba, Japan, [3]Haskins Laboratories, CT, USA

## ABSTRACT

In order to complete a computational model of speech production incorporating the relations between neuromotor, biomechanical and aerodynamic levels of description, a means for assigning linguistically relevant input strings is needed. This paper describes a two-phase curve-fitting and template matching method for mapping between phonemes and articulator movement behavior based on via point estimation. Among other things via point estimation provides a good means of data compression and the estimation software has proved to be an excellent empirical tool.

## INTRODUCTION

The goal of our computational approach to speech production has been to model mappings between conventional cognitive, neuromotor, and biomechanical levels of description for skilled motor behavior [5, 10]. In our conception, the act of generating speech is presumed to begin with an intention to speak. In order for this intention to be realized behaviorally, it must be encoded and mapped to a motor plan for neuromuscular activation. The speech motor plan, whether continuous or discrete, whether symbolically represented or not, is executed as a set of time-varying forces affecting vocal tract shape and the condition of sound source mechanisms. Aerodynamic principles govern the relation between physical characteristics of the vocal production system and the resulting speech acoustics. Thus, the mappings necessary to characterize speech production may be summarized as (1) the mapping between linguistic intent and an executable neuromotor plan, (2) the biomechanical mapping between neuromuscular activity and the changes of articulator configuration, and (3) the mapping between articulator configuration/vocal tract shape and the acoustics. These stages are shown in Figure 1.

Artificial neural networks have been used to model the latter two mappings: the dynamics relating muscle activity and articulator motion; and the forward mapping between articulator motion and PARCOR coefficients of the resulting acoustics for both reiterant and real speech [3]. After network training, recognizable speech acoustics are generated from continuous input of muscle activation (EMG) signals.

Perhaps the most challenging aspect of the model remaining to be implemented, and the subject of this paper, is the first mapping between cognitive intentions to act and their neuromotor implementation. As a starting point, it was hypothesized that linguistic intentions to speak are implemented serially as phoneme-specific sequences of *via point* targets in articulator task space [6]. The notion of the via point is based on curve-fitting procedures in which one or more intermediate (via) points are needed to fit a spline-function to non-straight lines between two endpoints. In biological and robotic motion control, via points serve as parameters for some kinematic or dynamic optimization function (e.g., minimum jerk [2], minimum torque or motor command change [1, 5, 8]) that constrains the possible trajectories which can be drawn between two endpoints. In our current formulation, phoneme-specific via points are used to guide the generation of neuromotor command sequences.

For the relatively simple case of reiterant speech using *ba*, there were only two phonemes and their articulation could be reasonably specified within the same articulator group composed of the jaw and two lips. One x-y coordinate value was chosen for the vowel and consonant based on the observed maximum and minimum positions, respectively, of lip aperture — a task variable derived from the contributions of the upper and lower lips and the jaw [4]. Since /b/ and /a/ were in strict alternation, via point timing was derived from a simple harmonic oscillator, whose frequency was the syllable rate [9]. Within each clause, consonant and vowel via points were placed at equal intervals, as shown in Figure 2.

This scheme for via point assignment worked very well for *ba*, whose production can be specified at normal and fast speech rates by a single task variable made up of the jaw and lips. Despite the isochronous timing and singular spatial assignment of the via points, the typical asymmetries for raising and lowering gestures and stress-related duration-amplitude modulation were produced (for discussion, see [11]). The network that estimated the EMG necessary to move the articulators toward the via point targets incorpo-

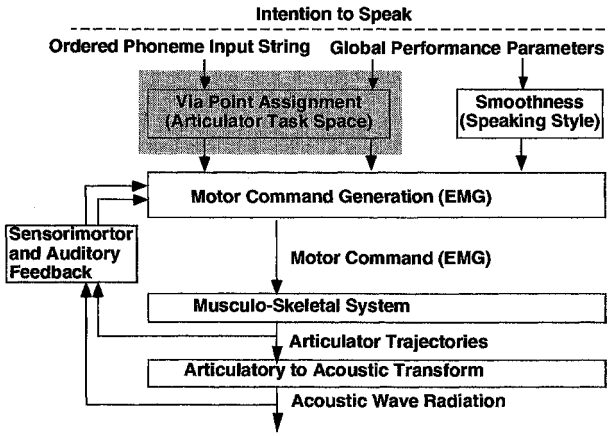


*Figure 1. Overview of Speech Production model. Shading denotes the topic of this paper.*

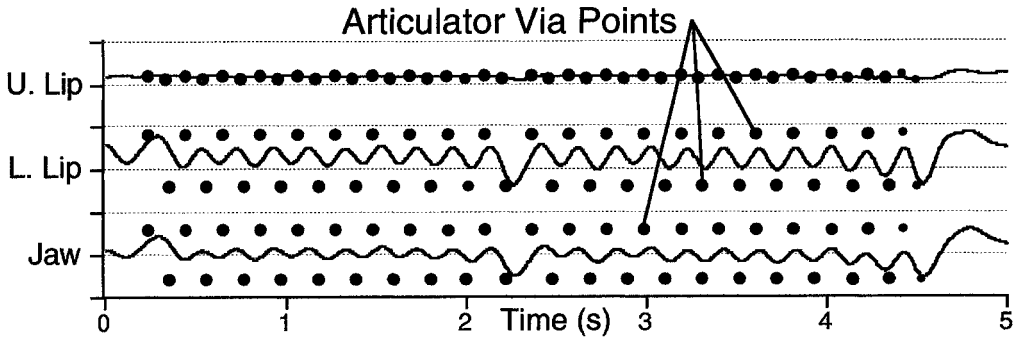


*Figure 2. Articulator via point assignment for a reiterant speech production of* When the sunight strikes raindrops in the air, they act like a prism and form a rainbow, *using* ba. *Shown are vertical position for the jaw and upper and lower lips. Via points for /b/ are typically above each trace and those for /a/ are below.*

rated a smoothness constraint which prevented output trajectories from achieving the via point targets and resulted in appropriate coarticulation.

Unfortunately, such a simple method for via point assignment will not work for real speech. In particular, phoneme-specific articulator behavior is generally not in strict alternation within a single articulator complex. Instead, there will be varying amounts of co-production among different articulator groups (e.g., lip-jaw, tongue-jaw). Even if we could identify primary articulator specifications for

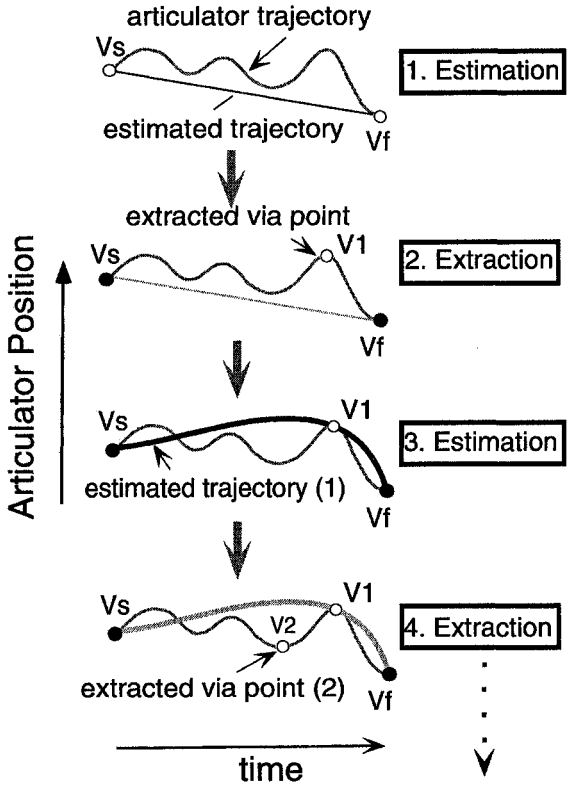


*Figure 3. Steps of iterative via point assignment by minimum jerk (MJ) estimation. 1. Minimum jerk is estimated for phrase or sentence between known start ($V_0$) and end ($V_f$) points. 2. If the summed error threshold (S) for all articulator trajectories (e.g., jaw + tongue tip+tongue blade + lower lip) is reached and if the error threshold for one or more specific articulators (e.g., $E_{jaw}$) is reached, a via point ($V_1$) is assigned to the data point at the maximum distance from any MJ trajectory and to the same data point for all other articulators satisfying the error criterion. 3. A new MJ trajectory is calculated through $V_0$, V1, and $V_f$. 4. The next via point (V2) is assigned by repeating 2. Steps 3-4 are repeated until S cannot be reached.*

phoneme production, secondary articulations such as upper lip positioning during /s/ production are ubiquitous. Indeed, it is unlikely that any articulator behavior is truly free to vary without consequence on the acoustic output; so, mapping phonemes to articulation could effectively incorporate the entire vocal tract. Therefore, before the via point hypothesis can be tested with real speech, a means must be found for segmenting the continuous articulator behavior into discrete associations.

In this paper, we describe a two-phase method for automatically assigning phoneme labels to articulator streams using via point estimation to define phoneme-specific templates (training phase) followed by segmentation of test sentences via template matching (recognition phase). The method was originally developed as part of a dynamics-based model for cursive handwriting recognition [12]. The bulk of the paper focuses on the training phase. In particular, the software we have developed for extracting via points has proved to be a far more valuable tool than was originally imagined. In addition to allowing us to address empirically a variety of problems associated with the segmentation of continuous, multidimensional behavior, the method provides the means to examine interarticulator coordination and invariance, and to quantify compact features for articulatory synthesis.

## EXTRACTION OF VIA POINT TEMPLATES

### Basic Method for Via Point Estimation

For the training phase, our goal has been to construct a codebook of phoneme-specific templates extracted from a corpus of approximately 3200 spontaneously produced sentences, which will provide templates for all English diphone and triphone combinations in all applicable word positions (e.g., initial, medial, final). Articulatory measures include sagittal plane position of the jaw, lips, and 4 points on the tongue recorded with a magnetometer (EMMA [7]). Via points for the entire sentence (defined at acoustic onset and offset) are assigned to the articulator movement trajectories by iteratively applying a 5th-order spline function, which corresponds to minimum jerk estimation. The steps of this process are shown in Figure 3 and have been summarized previously [10, 12].

Once via point estimation for the sentence is completed, phoneme-specific exemplars are then identified and extracted. Typically, each template contains from 0-4 via points per articulator channel. Figure 4 exemplifies the via point estimation and template extraction process for an exemplar of /k/. As shown in the figure, even when using fairly strict error criteria from which a very good fit is obtained, the set of estimated via points is quite small (only

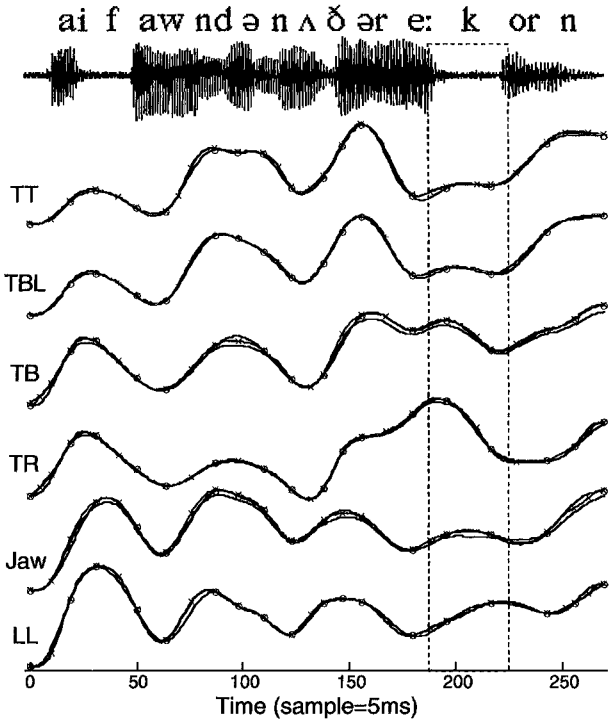


*Figure 4. Observed trajectories for the vertical position of 6 articulator measures are compared with those generated by via point estimation for two summed error criteria — S: 0.06, 0.006. Via points for each error criterion are denoted by circles (0.06) and x's (0.006). The 6 measures are: tongue tip (TT), blade (TBL), body (TB) and rear (TR); lower lip (LL); and jaw. The boxed area corresponds to the template extraction region for /k/ (see Figure 5). The utterance is* I found another acorn.

about twice the number of phonetic segments). The resulting data compression is more than 10 to 1. If the error criterion is relaxed, e.g., by an order of magnitude, the number of via points is almost halved and the data compression doubled, though at some cost to the quality of the fit (also see Figure 5).

### Via Point Estimation Options

The via point estimation software employs the algorithm developed by Wada and Kawato [13] and a graphics interface (written by Mark Tiede) that facilitates manipulation of various display and analytic parameters. Currently, the software is run on a DEC ALPHA, but can be recompiled for any computer platform supporting MOTIF. The principle analytic degrees-of-freedom affecting via point estimation are: 1. the boundaries and size of the analysis window, 2. the value of the summed error criterion (S), 3. the choice of articulator channels and the channel-specific error weightings, and 4. the total number of iterations to be performed. The effects of each option are discussed below.

**The analysis window.** Ideally, the endpoints of the analysis window should be at points of zero velocity and acceleration, because the initial minimum jerk estimation assumes this. With multi-channel data for real speech, this is not practical because the articulators do not start from synchronous, static pre-speech postures. Although the distortion introduced by potentially moving endpoints needs to be assessed, it is effectively minimized when the window size is the length of a sentence and the via points of interest correspond to a syllable or two somewhere in the middle.

**The summed error criterion (S).** Choice of error criterion affects the number of via points and the quality of the curve fit. Figure 5 shows the effects of different error criteria on the portions of the articulator trajectories associated with acoustic /k/. There are two conflicting requirements in choosing appropriate error criteria: strict error criteria produce good fits of the data — i.e., more via points (*x*'s); more relaxed criteria produce poorer fits, but which may generalize better to other utterances. In constructing the codebook, a balance between these requirements must be found. Clearly, this is an empirical problem, as are all of the options discussed in this section, and not one that can be easily be solved by comparing small test sets.

**The choice of channels and articulator weights.** Similar to summed error effects, but with different implications, is the distribution of articulator specific weights. Although we have found no reasonable metric for determining *a priori* the relative importance of an articulator to the production of a particular phoneme, e.g., distinctions between primary and secondary articulators, it is unlikely that equal articulator weightings should be assigned in all cases. For example, tongue motion associated with flanking vowel production can alter the order and location of via point assignments for a bilabial stop — e.g., *abi*. Even though tongue behavior in such a case may have some influence on /b/ production, it perhaps should not be given as much weight as the lips and jaw.

Another concern is that the different articulators have very different physical characteristics, which might introduce subtle distinctions to their movement trajectories. For, example, the jaw is a rigid skeletal structure, while the tongue is highly deformable, and both have far more mass than the lips. Since the largest and fastest motions will tend to guide via point assignment, slower changes in position as seen for the jaw or upper lip may not affect the analysis as much as

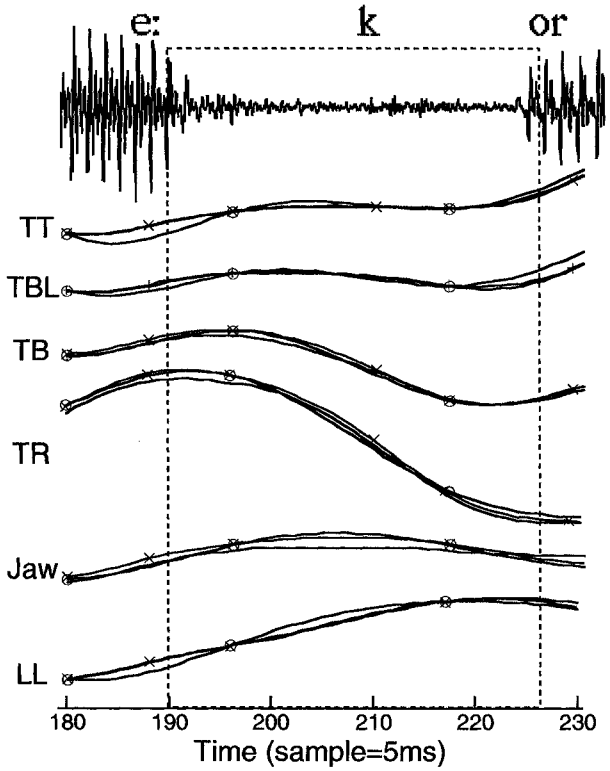


*Figure 5. Detailed comparison of observed and estimated trajectories at two error levels (S: 0.06, 0.006). The dashed box denotes the region of template extraction for the /k/ in* acorn. *Note there are fewer via points (circles) and the trajectory estimate is not as good for the 0.06 error.*

they should even though they may be critical to properly distinguishing among phonemes.

Finally, differences in the set of articulators chosen for analysis will affect via point estimation and subsequent recognition. That is, an articulator's presence or absence in the via point template will hamper or facilitate distinctiveness among exemplars. While this can be a problem for the analysis, the ability to choose different articulator subsets does provide an empirical means for examining an articulator's relative contribution to via point estimation for specific phonemes.

**Limits on the number of iterations.** This feature was added to make possible more systematic comparison of curve fits at different error rates. Since we may want to limit the number of via points to some unit average (e.g., syllable, segment), such comparison allows optimum error rates to be determined.

## PHONEME LABELING

Despite the benefits of the training phase for empirical analysis of highly compressed, multi-channel articulatory data, the primary goal of this research is to achieve automated phoneme labeling. In the recognition phase, the trajectories for new sentences are matched with the stored templates, resulting in a labeled phoneme string. The algorithm [12] iteratively calculates fits for different templates for a range of temporal windows and window positions after spatiotemporal normalization of both the test trajectories and the phoneme-specific templates [13]. Although there is nothing in principle to prevent fitted templates for successive phonemes from complete temporal overlap, there is currently a limit on the number of anticipatory window shifts (starting from the end of the preceding template fit) that the algorithm will attempt.

Clearly, the success of the recognition phase depends on constructing a template codebook that generalizes to the test data without loss of phoneme-specific distinctiveness. In addition to trial and error using the via point estimation options discussed above, we want to derive canonical phoneme templates from the sets of allophonic exemplars. If successful, this will greatly reduce codebook size. Although we cannot expect the canonical sets to hold for different speakers, the process may reveal common constraints between canonical and allophonic sets, which may further inform the via point estimation phase.

Finally, although we have avoided *a priori* assumptions as much as possible, we have experimented with ways of reducing recognition time. One of the simplest and most effective means we have found is to subdivide the codebook into consonants and vowels, which are by and large clearly distinguished by whether they open or close the vocal tract. With very few exceptions consonants tend toward a closure in the tract relative to adjacent vowels. Therefore, the signs of the articulator's vertical velocities are checked at onset of the current test window; if positive, then template matching is attempted first for the consonant portion of the codebook; if negative, then the vowels are tested first. In the future, we hope to extend this simple taxonomy to place of articulation changes in the horizontal axis.

## SUMMARY

Although our basic conception of via points remains one of discrete sequences of phoneme-specific target values for guiding a motor command (EMG) estimation network, in this paper we have addressed them in more broadly defined terms. At one level of generality, via point estimation offers a potential solution to the longstanding problem of segmenting continuous multidimensional behavior. Even more generally, high-order spline fits of the sort described here provide an excellent means of compressing complex data into manageable arrays for machine analysis.

## ACKNOWLEDGMENT

We thank Kevin Munhall for patient reading and discussion of the manuscript, and Masaaki Honda for cogent criticism of the model.

## REFERENCES

[1] Dornay, M., Uno, Y., Kawato, M., & Suzuki, R. (1992). Simulation of optimal Movements using the minimum-muscle-tension-change model. In R. P. Lippmann, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in Neural Information Processing Systems 4* . San Mateo, CA: Morgan Kaufmann Publishers.

[2] Flash, T., & Hogan, N. (1985). The Coordination of Arm Movements: An Experimentally Confirmed Mathematical Model. *Journal of Neuroscience*, **5**, 1688-1703.

[3] Hirayama, M., Vatikiotis-Bateson, E., & Kawato, M. (1993). Physiologically based speech synthesis using neural networks. *IEICE Transactions*, **E76-A**, 1898-1910.

[4] Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Honda, K. (1992a). Neural network modeling of speech motor control. In *Proceedings of the International Conference on Spoken Language Processing-1992,* **2** (pp. 883-886) Banff, Canada:

[5] Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Jordan, M. (1992). Forward dynamics modeling of speech motor control using physiological data. In R. P. Lippmann, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing systems 4* (pp. 191-198). San Mateo, CA: Morgan Kaufmann Publishers.

[6] Kawato, M., (1989). Motor theory of speech perception revisited from minimum torque-change neural network model. *Proceedings of the 8th Symposium on Future Electron Devices.* 141-150.

[7] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992). Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. *JASA*, **92**, 3078-3096.

[8] Uno, Y., Kawato, M., & Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement — minimum torque-change model. *Biol. Cyber.*, **61**, 89-101.

[9] Vatikiotis-Bateson, E., Hirayama, M., Honda, K., & Kawato, M. (1992). The articulatory dynamics of running speech: Gestures from phonemes? In *The International Conference on Spoken Language Processing-1992*, **2** (pp. 887-890). Banff, Canada:

[10] Vatikiotis-Bateson, E., Hirayama, M., Wada, Y., & Kawato, M. (1993). Generating articulator motion from muscle activity using artificial neural networks. *Annual Bulletin of RILP* (Univ. of Tokyo), **27**, 67-77.

[11] Vatikiotis-Bateson, E., & Kelso, J. A. S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *Jour. Phon.*, **21**(No. 3), 231-265.

[12] Wada, Y., & Kawato, M. (1993). A neural network model for arm trajectory formation using forward and inverse dynamics models. *Neural Networks*, **6**, 919-932.

[13] Wada, Y., Koike, Y., Vatikiotis-Bateson, E., & Kawato, M. (1994). A computational model for cursive handwriting based on the minimization principle. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in Neural Information Processing Systems 6* (pp. 727-734). San Mateo, CA: Morgan Kaufmann Publishers.