Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona, Spain.
August 3–9, 2003. Universitat Autònoma de Barcelona.

1299

CASY: The Haskins Configurable Articulatory Synthesizer

Khalil Iskarous[†], Louis M. Goldstein[†a], D.H. Whalen[†], Mark K. Tiede[†b] and Philip E. Rubin[†c]
† Haskins Laboratories, New Haven, USA
[a] Yale University, [b] MIT, [c] Yale University School of Medicine
E-mail: iskarous@haskins.yale.edu

ABSTRACT

In this paper, we outline some recent developments in the Haskins configurable articulatory synthesizer, CASY. The main advances are the adoption of a conic section model of the tongue body, extension of the configurability of the synthesizer parameters to model speaker variation, and the adoption of a new configurable area function model. These developments allow the system to accurately model the static and dynamic properties of the articulatory process, as well as the natural biological variation in the details of the process.

1. INTRODUCTION

As with many speech synthesis techniques, a principal goal for articulatory synthesis systems is to produce acoustic signals that resemble human speech with high accuracy. Perceptual judgment then determines the quality of the system. However, speech perception is heavily influenced by the phonetic and phonological knowledge of the perceiver [1]. Therefore, parameters of the acoustic signal may be unnatural, but judged to be acceptable, since the perceiver employs a phonetic specialization that is acutely sensitive to phonetic information, but quite flexible about other aspects of the signal. Furthermore, speech perception compensates for coarticulation [2], indicating that dynamic aspects of the signal are more important than the occasional steady-state portions of the signal. The goal of the Haskins articulatory synthesis program is to base the synthesis on articulatory motions measured by different speech production tracking and imaging systems such as EMMA and ultrasound. Accurate reproduction of articulatory motions combined with high quality mappings from articulator positions to area-function and from area function to sound leads to natural variation in the acoustic parameters. In this paper we present recent developments to the Haskins articulatory synthesizer CASY (Configurable Articulatory Synthesizer). In section 2, we present criteria that have guided the development of CASY. Section 3 presents a new model of the body of the tongue using conic arcs more general than the circle, as well as applications of this model to matching static and dynamic speech data.

2. CRITERIA FOR ARTICULATORY SYNTHESIS

To achieve the goal of accurate reproduction of articulatory motions while capturing inter-subject variability, we have identified the following four criteria for the development of the system:

a) Accurate Static Configurations: The articulatory synthesis system must produce static configurations of the articulators that closely match observed configurations for different speakers. This does not mean that the system must be able to reproduce every anatomical nuance in every individual vocal tract. Rather the goal is to parameterize articulatory configurations with as few parameters as possible, but still be able to capture commonly seen configurational variability. This criterion may seem too trivial to state, but the link between articulatory synthesis and speech production measurement has not been strong. This is the case even in factor-analytic approaches to speech synthesis, where the factors are indeed derived from speech production data. Any vocal tract shape in the dataset from which the factors are derived can of course be reproduced as a linear combination of the factors with great accuracy, therefore such systems are able to match real speech production data. However, not every linear combination of the factors corresponds to an actual or potential vocal tract shape, therefore the articulatory output of such systems can contain unattested vocal tract shapes. These systems are also limited by strong dependency on the initial data used for deriving the factors.

b) Accurate Dynamic Motions: The system must also be able to accurately reproduce the intermediate configurations of the articulators between two phonetic targets. This is very important, since inspection of natural language spectrograms shows that most of a speech utterance consists of dynamic variation rather than steady states. Even if one formant is constant for a few tens of milliseconds, other formants are usually changing. Accurate reproduction of articulator dynamics for lips and velum are relatively straightforward, since at least linguistically these are low-dimensional articulators. Tongue dynamics is a great deal more difficult to reproduce, since the tongue is a high-dimensional articulator with several loosely connected parts. In addition, different parts of the tongue can simultaneously carry out different phonetic tasks for consecutive segments, so the tongue model in the synthesizer must be able to match this complexity. An important goal in dynamic synthesis is to identify commonly occurring patterns of tongue motion from speech imaging data and then choose the articulatory synthesis parameters in such a way as to be able to model the dynamics.

ISBN 1-876346-48-5

c) Configurability: The control parameters of the articulatory model must be configurable enough to match the natural biological variation of human vocal tracts. Speaker variation is of course also present in the acoustic signal, but the variation at that level is a consequence of variation in vocal tract sizes and peculiarities. Also, variation in one articulatory parameter can simultaneously affect several acoustic parameters across different segments, e.g. lip shape will affect vowel formants, fricative noise spectra, and stop burst spectra. Capturing speaker variability with articulatory parameters is therefore more parsimonious than capturing it with acoustic parameters.

d) Lingustic Control of Parameters: The control parameters of the model need to be as close as possible to the linguistic characterization of the segment, e.g. we should be able to specify place of articulation and degree of articulation in as direct a manner as possible. The issue here is the tradeoff between the geometric and linguistic characterization of the articulatory synthesis parameters. Accurate geometric modeling of the shapes of speech organs requires parameters like radius and location of center of circle for characterizing tongue shape, however linguistic specification of a vowel, for instance, is usually done in terms of its articulatory features like height and backness. We attempt to resolve this tradeoff by choosing parameters that are geometrically simple, yet linguistically meaningful.

3. GEOMETRIC PARAMETERS OF TONGUE SHAPE

Articulatory Synthesis (ASY) at Haskins began in the 70' s [3]. The system was based on work by Mermelstein, Coker, and Fujimura [4,5] on representing the speech articulators using simple geometric parameters like angles, line segments, and circles. The tongue body was represented as the arc of a circle whose center and radius could be varied. A later innovation (CASY) was to focus on the *configurability* of the synthesizer parameters to account for speaker-variation by adjusting parameters to fit MR images of the vocal tract [6]. Figure 1 shows the tongue as represented by CASY.

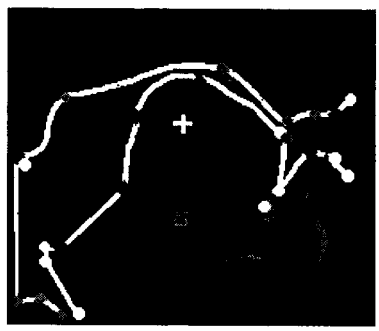

Figure 1: CASY vocal tract outline during [i] superimposed on an MR image.

The shape of the tongue for several vowels and consonants for some individuals is well represented using the circle model of the tongue. Figure 2 shows the tongue as seen through ultrasound imaging for [e] and [ə], together with circle fits of the tongue body.

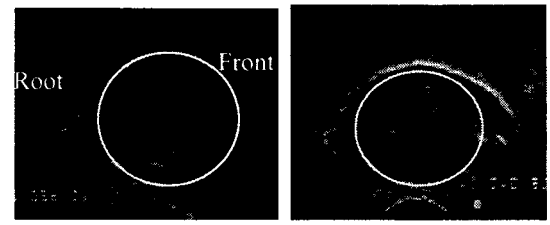

Figure 2: Ultrasound images of [e] and [ə], together with a circle fit of the body of the tongue.

Indeed by varying the radius and center of a circle, one can model a large number of tongue shapes, however there are many tongue shapes in which the area of the tongue that would be modeled by the arc of a circle is simply not shaped like a circle arc. A circle arc has uniform curvature, whereas many tongue shapes have a curvature that varies appreciably through the tongue body. Figure 3 shows ultrasound images for [r] and [ʃ], where curvature of tongue body is different from the uniform curvature of the circle. Of course a small enough circle can model a small portion of the tongue body, where curvature is uniform, but it would be advantageous to find geometric curves like the circle with a small number of controllable parameters, but that are non-uniform in curvature and can therefore fit a larger portion of the tongue.

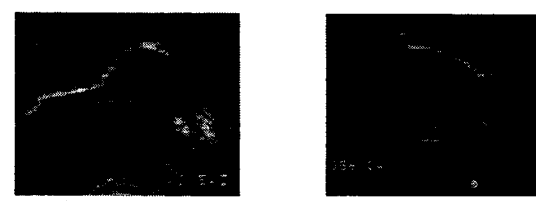

Figure 3: Ultrasound images of tongue for American English [r] and [ʃ].

Another problem is that even though the variation in the circle center location can indicate height and backness, it is not easy to translate between this geometric parameter and linguistic control parameters of tongue shape like degree and place of articulation. We have therefore sought a geometric model of the tongue body that allows for nonuniform curvature, and whose control parameters have a clearer linguistic meaning.

In the current version of CASY, the circle model of the tongue is generalized to a conic arc model. This new model allows a richer set of shapes for the body of the tongue, including elliptic, hyperbolic, and parabolic arcs, as well as the circular as a special case. The major improvement is that the conics other than the circle have non-uniform curvature, as can be seen in Figure 4, where the conics are obtained by intersecting a double cone with a plane. The leftmost panel shows the circle with its uniform curvature.

ISBN 1-876346-48-5 © 2003 UAB

Non-uniform curvature increases as one goes from ellipse to parabola, and finally to the hyperbola which shows extreme variation in curvature, with one region being very highly curved and another approximating a line (zero curvature). Since curvature varies along the tongue body, we can use conic arcs to approximate a large portion of the tongue. Using a circle limits us to a small portion where the curvature is uniform.

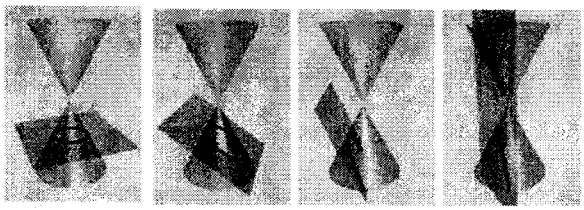

Figure 4: Conic sections as the intersection of a double cone and a plane. From left: circle, ellipse, parabola, hyperbola.

In Euclidean geometry, the different conics are qualitatively different curves that require different parameters for their specification. In projective geometry, on the other hand, one can choose two parameters which when continuously varied, will yield all the conics [7]. Figure 5a shows continuous variation in one of the parameters, which we will call curvature, since variation in that parameter takes us from a line (zero curvature) to elliptic shapes (only slightly curved), a parabolic shape, and finally to hyperbolic shapes (highly peaked). Even though the parameter is called *curvature*, however, it must be noticed that the curvature is non-uniform—*curvature* refers to how peaked the most peaked portion of the conic arc is. The second parameter that can be varied is the *tiltof* the conic arc. Figure 5b shows two families of conics with two different tilts. When we continuously vary these two parameters, therefore, we get an infinite variety of shapes that can model the tongue.

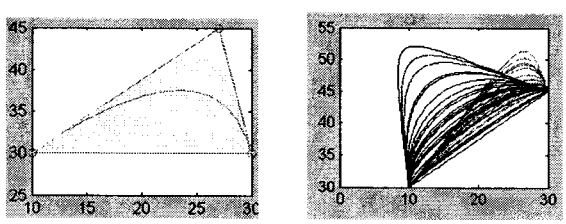

Figure 5: a) Continuous variation in the curvature parameters. **b)** Continuous variation in the curvature parameter for 2 values of the tilt parameter.

What makes this parameterization of the conics especially useful for articulatory synthesis is that these purely geometric parameters can be given a linguistic interpretation: the *curvature* parameter represents the degree of closure and the *tilt* parameter represents the place of articulation within the vocal tract. As Wood has shown, both vowels and consonants can be classified using these parameters [8]. Conic sections thus represent a parametrization of tongue shape that is both geometrically natural and linguistically natural.

The fact that low level descriptors of tongue shape also have a high level meaning as segmental features allows us to derive parameters for synthesis directly from speech production data. We have developed an optimization algorithm that derives the best parameters for the tongue edge as imaged by an ultrasound system. Figure 6 shows the fit to segments. Derivation of the synthesis parameters from articulatory data automatically enforces a close match between CASY articulatory output and natural articulatory data. For tongue segments like American English [r], there are two constrictions and each is modeled with a separate conic. Such shapes could not be generated by the original circle model of the tongue.

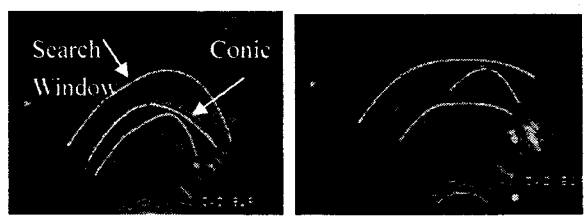

Figure 6: a) Conic fit for entire tongue for [ə]. **b)** 3 Conic fit for American English [r].

As stated earlier, the key to articulatory synthesis is to properly model the transition from the target (coarticulated or not) of one segment to that of another. Speech synthesis systems that work directly with acoustic parameters use simple linear or nonlinear interpolation to model acoustic dynamics. The result is speech that is comprehensible, but somewhat unnatural.

Since acoustic dynamics is a consequence of articulatory dynamics, through a nonlinear relation, it is possible to model dynamics more directly and more parsimoniously through the articulatory parameters. The idea is to parameterize articulatory dynamics using as few parameters as possible and to use as simple a dynamic rule as possible for the parameters, while at the same time capturing known patterns of natural articulator movements as evidenced in speech production data.

On the face of it, however, the description of articulatory dynamics is a very complex problem. It has taken a great deal of time for researchers to establish characteristic shapes of the articulators for the targets of the vowels and consonants of a phonetically well-studied like American English—and the description of the statics of vowels and consonants is not yet complete. To describe American English articulatory dynamics, one would have to know how each articulator moves from each segmental target to every other target. This is a larger problem, since the number of possible transitions is much larger than the number of segments.

It was shown by one of the authors [9], however, that at least for the tongue (arguably the most biologically complex and acoustically influential articulator), there are

ISBN 1-876346-48-5 © 2003 UAB

two highly characteristic patterns that are reused in a large number of transitions—the pivot and the arch. Both can be characterized as maximizing tongue motions orthogonal to the longitudinal axis of the vocal tract at sections of the vocal tract where there are target locations, and minimizing such motions between such targets—at the pivot points, only motions tangential to the longitudinal axis are allowed. Figure 7 shows a few examples of the pivot pattern. This allows for a very straightforward parameterization of articulatory dynamics in terms of the same conic parameters used earlier. A pivot can be accomplished by simultaneous change in the curvature and tilt parameters. The same parameters therefore accomplish high fidelity to natural segmental statics and dynamics.

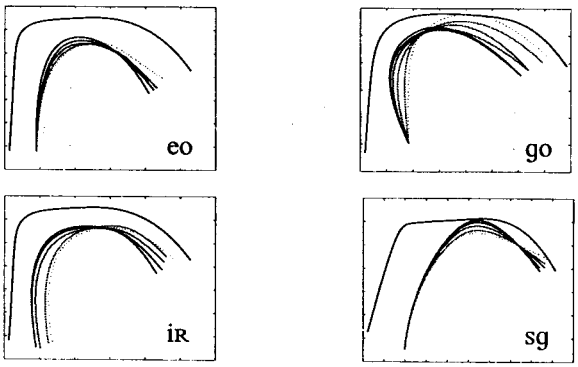

Figure 7: Pivots for [eo], [go], [iR], [sg].

Another improvement in CASY is the inclusion of modifications to the distance-to-area functions to more closely conform to individual differences. The original functions derive from the alpha/beta model of Heinz and Stevens [10]. While the mathematical formulation itself has continued to prove reliable [11], it is now possible to obtain better area data from MRI [12]. Use of these values has been shown to improve the match between synthesized formant patterns and natural ones [11], which will make articulatory synthesis more accurate for those speakers with substantial MRI data available. We will begin evaluating whether more generic versions (matching only sex and vocal tract length, for example) are equally useful. These measures should make marked improvements to articulatory synthesis, though clearly more improvements could be made by using a three dimensional model of the tongue instead of the current two dimensional one.

4. CONCLUSIONS

We have outlined recent developments to the CASY system, including generalization of the tongue body model from a circle to a conic. The advantage of the conic parameters is that they have a low level interpretation in terms of tongue geometry, as well as a high level interpretation as segmental features.

ACKNOWLEDGMENTS

This work was supported by NIH grant HD-01994 and DC-02717 to Haskins Laboratories.

REFERENCES

[1] C. Best, "A direct realist view of cross-language speech perception," In Winifred Strange (ed.), *Speech perception and linguistic experience: Theoretical and methodological issues.* Baltimore: York Press. 171–204, 1995.

[2] A. M. Liberman and D. H. Whalen, "On the relation of speech to language," *Trends in Cognitive Sciences,* vol. 4,187-196, 2000.

[3] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *Journal of the Acoustical Society of America,* vol. 70, pp. 321-328, 1981.

[4] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America,* vol. 53, pp. 1070-1082, 1973.

[5] C. Coker and O. Fujimura, "Model for specification of the vocal-tract area function," *Journal of the Acoustical Society of America,* vol. 40, p. 1271, 1966.

[6] P. Rubin, E. Saltzman, L. Goldstein, R. McGowan, M. Tiede, and C. Browman, "CASY and extensions to the task-dynamic model," In Proceedings of the 1st ESCA ETRW on Speech Production Modeling—4th Speech Production Seminar, Autrans, 1996.

[7] K. Iskarous, D.H. Whalen, and I. Mattingly, "Modeling tongue shapes with conic arcs," *Journal of the Acoustical Society of America,* vol. 110, p. 2760, 2001.

[8] S. Wood, *X-ray and model studies of vowel articulation.* Working Papers 23. Lund, Sweden: Dept. of Linguistics, Lund University, 1982.

[9] K. Iskarous, *Dynamic Acoustic-Articulatory Relations,* University of Illinois at Urbana-Champaign Doctoral Dissertation, 2001.

[10] J. Heinz and K. Stevens, "On the derivation of area functions and acoustic spectra from cineradiographic films of speech," *Journal of the Acoustical society of America,* vol. 36, p. 1037, 1964.

[11] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI," *Speech Communication,* vol. 36, 169-180, 2002.

[12] T. Baer, J. Gore, S. Boyce, and P. Nye, "Application of MRI to the analysis of speech production," *Magnetic Resonance Imaging,* vol. 5, pp. 1-7, 1987.

ISBN 1-876346-48-5