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

1315

Articulatory synthesis: Advances and prospects

D. H. Whalen†
† Haskins Laboratories

E-mail: whalen@haskins.yale.edu

ABSTRACT

Speech synthesis has achieved a high level of quality and usefulness, but the flexibility of diphone and other acoustically based schemes is limited. Articulatory synthesis holds promise for overcoming some of the limitations and for sharpening our understanding of the production/perception link. Progress has been made on several aspects of articulatory synthesis, as the papers in this session demonstrate. Better measures of articulation and acoustics have led to improvements in the models. Simpler control structures have been proposed for the tongue. In addition, the inclusion of visual synthesis has become a priority.

1. INTRODUCTION

The synthesis of speech from simple numerical parameters has been a great achievement in the field of speech science. After the initial efforts in creating artificial talking machines in the 17th [1], 18th [2] and 19th (Faber's machine, described in [3]) centuries, we entered the latter half of the 20th century with a focus on formant synthesis. This was based on the success of the sound spectrograph at analyzing natural speech in terms of its spectrum [4] and the perceptibility of aspects of that spectrum [5, 6]. The development of manipulable formant synthesizers allowed the creation of virtually any speech sound [7, 8]. Once it became apparent that such relationships could be relatively well specified by rule [9], the need for control parameters that resembled actual articulation diminished. These days it is possible to construct diphone synthesis programs for most languages in a short period of time [10]. The general identifiability of the synthesis is so high that further improvements can only be measured with techniques that emphasize adverse conditions (noise, memory load) or time pressure (reaction time, long texts) [11]. The overall quality, especially for long stretches of speech, has not improved as much. As Henton [12: 117] has stated, "High quality, yet flexible, synthetic speech has been slow to appear." One reason that text-to-speech (TTS) has not become ubiquitous (despite years of such predictions) is the difficult of getting to this higher level of quality.

Improvements in TTS systems are likely to come from many different areas, since there are many different aspects that are in need of improvement [13]. The global information structure of an utterance, for example, may influence the way that phrasing and intonation are realized, and adding such features to TTS systems will be a very long term project. A wide variety of intonational issues need to be addressed, and these will also require a great deal of work since we do not well understand the behavior of natural speech prosody. The characteristics of the voice source are poorly modeled in current synthesis, and improvements there will also require a separate research program.

Improvements to the flexibility and segmental representation, however, may be accomplished via articulatory synthesis. The transitions between segments are one problem area for diphone systems [13], and articulatory synthesis would, with the proper control structures, eliminate those. In particular, it is possible that articulatory synthesis would restore the "pivots" that have been found to define most fluent transitions between segments [14]. The flexibility to change speaking rate and possibly emphatic stress could also be part of articulatory synthesis, again with the proper control structures. Certain adjustments due to the physical nature of the articulators (e.g., saturation effects) are more naturally handled here as well. The potential improvements are difficult to quantify, since the TTS systems are already at a high level of identifiability, but the gains for long-term usability of TTS could be great. Whether the segmental benefits alone would be enough to take TTS over the barrier between its current short-term usages and its use for longer stretches of speech remains to be seen. We are now in a position, though, to begin finding out.

Articulatory synthesis has been approached from several vantage points, all of which are represented in this symposium. Each has had good success in the essentials of speech synthesis, and yet each has fallen short of being found generally useful to the TTS world. Even as research tools, there have been limitations that have made articulatory synthesis less of a factor than one would expect for a field in which the acoustics analyzed are usually produced by a vocal tract, not an unconstrained series of formant resonators. The field is now reaching a point where the utility of articulatory synthesis for research and possibly for application can be addressed again.

One reason it is worth revisiting articulatory synthesis

ISBN 1-876346-48-5

now is that we have devised better ways of measuring the vocal tract than were available to the earlier researchers. These include the three-dimensional capability of MRI [15-17], the real-time tracking of flesh points with electromagnetometry [18, 19] and x-ray microbeam [20], and the imaging of the tongue surface via ultrasound [21]. With these new results, a more detailed version of the relationship between articulation and acoustics can be generated. The measurement of the articulators themselves is more complete as well. Greater accuracy in the midsaggital distance to area functions [22, 23] should result in more accurate synthesis; the important parameters remain to be tested.

Another reason for continuing to develop articulatory synthesis is an increased appreciation of the influence of the visual signal on speech perception. We have known for some time that showing the face producing the speech can give the equivalent of about a 12 dB gain for listening in noise tasks [24], but the visual signal can also supercede the acoustic signal in perception [25]. Two of the presentations in this symposium include the generation of a visual representation of the speech as an explicit goal of the synthesis (Fels et al., Badin et al.). The techniques for performing this visual synthesis are readily available now, in a way that they were not a decade ago. This holds promise for a wider range of applications that simply replacing parametric or diphone synthesis with articulatory synthesis.

The relationship between perception and production can be addressed more easily with articulatory synthesis than with formant synthesis. Changes in the kinematic pattern underlying a synthetic speech sound can be made to reflect what is actually present in the speech stream. Other changes can be made deliberately unnatural, in order to explore the listener's sensitivity to such distortions. The control structures that go into articulatory synthesis are assumed to be closer to the control structures needed by phonology than are those in concatenative or acoustically based synthesis [26], and thus phonological hypotheses should be more directly testable with more fully developed articulatory synthesis.

The approaches to the control structure are varied as well. One approach is to perform a principal components analysis which gives factors that describe most of the tongue shape (Maeda, Fels et al.). The tongue shape during speech is quite complex, but the amount of variance accounted for in the PC analyses is impressive, and the resulting synthesis can be quite good. Another approach is to have clusters of neuromusculature that can be grouped together to reduce the degrees of freedom (Badin et al.). This allows for realistic vocal tract shapes without too large a cost in terms of control, and also supports the generation of visible speech articulators, with the addition of the influence of the unseen articulators on what is visible. Hanson and Stevens take a quasi-articulatory approach, "Hlsyn." It represents an attempt to do articulatory synthesis with limited control parameters, a relatively small set of which are likely to have an effect on the acoustic output for both vowels and consonants (sonorants and obstruents). In their paper, they concentrate on the articulatory/acoustic issues in manipulating the various parameters that are needed for the synthesis of consonants, particularly consonant sequences, with some discussion of aerodynamics and respiration. Finally, the Haskins CASY model [27] takes a simplified model of the tongue to generate the vocal tract shapes directly. The model can either be the previously instantiated circle model of Mermelstein, [28] or the shape described by a generalized conic section, as outlined in the paper by Iskarous et al. in this session. Individual vocal tracts can be reasonably modeled both in the midsagittal shapes and in the distance-to-area functions.

Articulatory synthesis has not received as much attention in recent years as it might have, but the convergence of various technologies has brought us to the point of making major breakthroughs. The raw material for testing the efficacy of the synthesis is much greater, and the lack of a clear means of improving current TTS systems gives us an added reason for exploring this alternate technique. These five papers will, hopefully, demonstrate that articulatory synthesis deserves renewed attention from speech scientists.

ACKNOWLEDGMENTS

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

REFERENCES

References

[1] Z. Fagyal. "Phonetics and speaking machines: On the mechanical simulation of human speech in the 17th century," *Historiographia Linguistica,* vol. 28, pp. 289-330, 2001.

[2] W. v. Kempelen. *Mechanismus der menschlichen Sprache : nebst der Beschreibung seiner sprechenden Maschine.* Vienna: J.V. Degen, 1791.

[3] H. Dudley and T. H. Tarnoczy. "The speaking machine of Wolfgang von Kempelen," *Journal of the Acoustical Society of America,* vol. 22, pp. 151-166, 1950.

[4] R. K. Potter, G. A. Kopp and H. G. Green. *Visible*

speech. New York: Van Nostrand, 1947.

[5] A. M. Liberman, P. Delattre and F. S. Cooper. "The role of selected stimulus-variables in the perception of the unvoiced stop consonants," *American Journal of Psychology,* vol. 65, pp. 497-516, 1952.

[6] A. M. Liberman. *Speech: A special code.* Cambridge, MA: MIT Press, 1996.

[7] G. Fant. "Acoustic analysis and synthesis of speech with applications to Swedish," *Ericsson Technics,* vol. 1, pp. 3-108, 1959.

[8] D. H. Klatt. "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America,* vol. 67, pp. 971-995, 1980.

[9] I. G. Mattingly. "Experimental methods for speech synthesis by rule," *IEEE Transactions on Audio and Electroacoustics,* vol. AU-16, pp. 198-202, 1968.

[10]R. Sproat. *Multilingual text-to-speech synthesis: The Bell Labs approach.* Dordrecht: Kluwer, 1998.

[11]D. Gibbon, R. Moore and R. Winski. *Handbook of standards and resources for spoken language systems.* Berlin: Mouton de Gruyter, 1997.

[12]C. Henton. "Challenges and rewards in using parametric or concatenative speech synthesis," *International Journal of Speech Technology,* vol. 5, pp. 117-131, 2002.

[13]E. A. M. Klabbers. *Segmental and prosodic improvements to speech generation.* Unpublished Ph.D. thesis. Eindhoven: Technische U., 2001.

[14]K. Iskarous. *Dynamic acoustic-articulatory relations.* Unpublished Ph.D. thesis. Urbana-Champaign: University of Illinois, 2001.

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

[16]S. S. Narayanan, A. A. Alwan and K. Haker. "An articulatory study of fricative consonants using magnetic resonance imaging," *Journal of the Acoustical Society of America,* vol. 98, pp. 1325-1347, 1995.

[17]P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth and C. Savariaux. "Three-dimensional articulatory modeling of tongue, lips and face, based on MRI and video images," *Journal of Phonetics,* vol. 30, pp. 533-553, 2002.

[18]J. S. Perkell, M. Cohen, M. Svirsky, M. Matthies, I. Garabieta and M. Jackson. "Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements," *Journal of the Acoustical Society of America,* vol. 92, pp. 3078-3096, 1992.

[19]P. Schönle, K. Grabe, P. Wenig, J. Hohne, J. Schrader and B. Conrad. "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language,* vol. 31, pp. 26-35, 1987.

[20]J. R. Westbury. *X-ray microbeam speech production database user's handbook:* Madison, WI: Waisman Center, University of Wisconsin, 1994.

[21]M. Stone and A. Lundberg. "Three-dimensional tongue surface shapes of English consonants and vowels," *Journal of the Acoustical Society of America,* vol. 99, pp. 3728-3737, 1996.

[22]D. Beautemps, P. Badin and G. Bailly. "Linear degrees of freedom in speech production: Analysis of cineradio- and labio-film data and articulatory-acoustic modeling," *Journal of the Acoustical Society of America,* vol. 109, pp. 2165-2180, 2001.

[23]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, pp. 169-180, 2002.

[24]W. H. Sumby and I. Pollack. "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America,* vol. 26, pp. 212-215, 1954.

[25]H. McGurk and J. MacDonald. "Hearing lips and seeing voices," *Nature,* vol. 264, pp. 746-748, 1976.

[26]C. P. Browman and L. Goldstein. "Articulatory phonology: An overview," *Phonetica,* vol. 49, pp. 155-180, 1992.

[27]P. E. 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 and 4th speech production seminar* (pp. 125-128. Autrans: ICP Grenoble, 1996.

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

ISBN 1-876346-48-5 © 2003 UAB