15th International Congress on Acoustics
Trondheim, Norway 26 - 30 June 1995

KNOWLEDGE FROM SPEECH PRODUCTION USED IN SPEECH TECHNOLOGY: ARTICULATORY SYNTHESIS

Richard S. McGowan
Haskins Laboratories, 270 Crown Street, New Haven, CT 06511, USA

INTRODUCTION

There appears to be a continuing trend toward incorporating knowledge of speech production into speech technology– text-to-speech synthesis (e.g. Parthasarthy & Coker, 1992; Bickley et al., 1994), low bit rate coding (see Schroeter & Sondhi, 1992), and automatic speech recognition (e.g. Shirai & Kobayashi, 1986; Rose et al., 1994). For automatic speech recognition using knowledge of the coordination of the vocal tract articulators and the resulting acoustics can reduce apparent token-to-token variability so that general pattern recognition algorithms have less work to do. Using articulatory representations in speech coding has the potential of greatly reducing bit rate because the articulators move relatively slowly and may be described by a few parameters by using an underlying dynamical model or by using simple curve fitting. Finally, text-to-speech synthesis can be improved using articulator control parameters, because the laws of physics can be used to produce the correct bundle of acoustic features with a comparatively limited parameterization– the acoustic output is constrained by the laws of physics. All these applications that depend on articulatory representation of speech production, can be grounded in what is called an articulatory synthesizer. An articulatory synthesizer is a device that produces speech output from a set of articulatory parameters (an articulatory representation). These devices are usually implemented in software on a digital computer.

The production of speech using an articulatory synthesizer (the "forward" mapping) can be divided into two major components: that of finding the mapping from the linguistic units to the articulatory movement, and that of finding the mapping from the articulatory movement to the aerodynamic state and acoustic output. The forward problem is solved with the composite mapping. The first mapping is the domain of people interested in the control of human movement and coordination as it relates to the vocal tract during speech, which includes some linguists and some experimental psychologists. The second mapping from articulatory movement to aerodynamic state and acoustics is the domain of acousticians. It is easily seen that both components of the forward mapping are important for the named technical applications. To perform text-to-speech synthesis from an articulatory point-of-view, the composite mapping is constructed, and to perform low bit rate coding or automatic speech recognition, one would need to find the inverse of the composite mapping, if the approach is to use articulatory information. If an analysis-by-synthesis procedure is used to construct the inverse composite mapping, then it is necessary to construct each component forward mapping.

TASK DYNAMICS

Only one example of one part of mapping from linguistic units to articulatory movement will be discussed here. This example is the model of articulatory coordination for articulators in performing speech gestures used at Haskins Laboratories, known as task dynamics (Saltzman & Munhall, 1989). This model describes the formation and breaking of constrictions in the vocal tract using a set of independent, linear, second-order differential equations: one equation for each constriction (e.g. labial, tongue body, or tongue tip). Because constrictions can be made with the coordinated activity of vocal tract articulators, such as lips, jaw and tongue, these equations are transformed into a model for articulator geometry. In the articulator coordinate system, the equations for constriction dynamics become coupled and nonlinear, and the pseudo-inverse of the Jacobian is used in their solution because there are more articulator than constriction degrees-of-freedom. This means that several articulators can be used to attain the same constriction target (upper lip, lower lip, or jaw can be used to attain lip closure).

Task dynamics models phenomena that are observed in real speech behavior. This includes *articulatory compensation*, where one articulator compensates for another that cannot move. (e.g. the lips can increase their total movement to close the mouth when the jaw cannot move.) The other pervasive phenomenon in speech that task dynamics models is that of *coarticulation*. For instance the jaw can be used to close the mouth or it can be used to lower the body of the tongue for certain vowels, such as /a/. When there is mouth closure, say for /b/, followed by an /a/ both goals influence the jaw, so that the mouth closure is probably attained by more lip involvement than would be without the presence of the /a/. This is so the jaw can be lower for the following /a/.

There are some real advantages to using task dynamics in the technical applications to be considered. The first is that constrictions of the vocal tract and the output acoustics are closely related so that the analysis-by-synthesis that recovers task dynamic parameters from speech is facilitated. Further, phonology based on articulatory gestures and instantiated in task dynamics is being constructed by linguists (Browman & Goldstein, 1990). This kind of work is necessary to finally map the task-dynamic parameters, such as the natural frequency of a lip closure, to linguistic units. This, of course, is required if automatic speech recognition is to be done using an articulatory representation.

Task dynamics takes the approach of finding the appropriate coordinate system to define speech behaviors (currently, constriction dynamics) and a means of transforming this coordinate system into a physical coordinate system (vocal tract articulators). This approach is extremely valuable in attempting the man-machine applications that are named above. However, there is room for an evolution in the details of this approach. For instance, the aerodynamics of the vocal tract appear to be controlled in a task specific way, and thus these must be included in some way (McGowan & Saltzman, in press). It is not clear whether all vocal tract gestures use constriction targets, and, in particular, vowels may need a more spatially global specification (Mattingly, 1990). Also, even where constriction targets are appropriate, there may be a region of targets rather than a point target (Guenther, 1994). The extension of this model should be undertaken to account for a variety of individual vocal tract shapes and for the sequencing of gestures, which is important for the rhythm mechanisms for rate and stress.

ARTICULATION-TO-ACOUSTICS

Where are we now in terms of the mapping from articulation to acoustics in articulatory synthesis, which is the second mapping that has to be constructed? What is the relation between the physics of fluid flow in the vocal tract and the propagation models that we are currently using? All the articulatory synthesizers known to the author use one-dimensional models of wave propagation (some with corrections for large area changes). The voice source is generated in a variety of ways, including simulations of self-oscillating vocal folds. The noise sources in the vocal tract are modeled as point sources, and their amplitude and frequency characteristics depend on aerodynamics in various degrees of sophistication.

Some synthesizers are time-domain synthesizers (e.g. Maeda, 1982), so that the waves created by the sources are propagated on a space-time grid. Other synthesizers use a frequency-domain transfer function to represent the wave propagation in the vocal tract (e.g. Sondhi & Schroeter, 1987; Davies et al., 1993). The output speech can be calculated by mapping the transfer function to the corresponding time-domain transfer function via an inverse discrete Fourier transform (DFT). An alternative is to find the poles and zeros of the transfer function and to use a formant synthesizer to produce the output speech (e.g. Lin, 1994). The remainder of the paper will suggest two research directions for articulation-to-acoustics mapping. The first is a proposal to use a set of orthonormal bases functions, other than circular functions, to represent the vocal tract transfer function. These bases functions are from what Coifman (1991, p. 881) calls a "library of wavelet packets", including wavelets, used in multiresolution analyses. The other proposal is to provide a four-parameter, articulatory model for the control of the voice source.

MULTIRESOLUTION SYNTHESIS

When a frequency domain transfer function is used to represent vocal tract wave propagation, a time domain transfer function is calculated as an inverse discrete Fourier transform (DFT), and the sources convolved with the resulting transfer function. To obtain reasonable frequncy resolution it is necessary that the transform window be of reasonable duration (25.6 ms for Sondhi & Schroeter [1987] for a 20kHz sampling rate). However, in performing the inverse transform, the vocal tract is assumed to be unchanging within the duration of the transform window; thus invoking the quasisteady (stationarity) approximation. There are speech environments for which this approximation may be inappropriate, including the closure and release of stops, fricatives, affricates, and approximants. Specifically, there are two possible problems in these speech environments. First, the filtering properties of the vocal tract may be rapidly varying, and, second, the source properties may be changing rapidly because of vocal tract changes. The voice source is affected by the configuration of the upper vocal tract in what is known as source-tract interaction. Also, the aerodynamic noise source properties of amplitude and spectral content are directly affected by the vocal tract configuration because of changes in constriction areas and pressure distributions. Thus, the quasisteady assumption is suspect in certain phonetic environments. In fact, this has been a problem in using DFTs for the analysis of speech.

A multiresolution decomposition may help in this regard (Meyer, 1993). In such a decomposition, the high-frequency components can be more localized in time than the low-frequency components. Thus, the high-frequency components can change more rapidly than the low-frequency components without violating stationarity. Recent work has been done in multiresolution decomposition and its generalizations for analysis and compression of speech signals (e.g. Wickerhauser, 1993). These decompositions make use of wavelets, wavelet packets, and other orthonormal bases to find decompositions suitable for a given application. In the case of data compression Shannon entropy can be minimized (Coifman & Wickerhauser, 1993). For purposes of speech analysis, the multiresolution decompositions allow the analyst to tile the time-frequency plane tailored to the physical situation. While there is still a trade between frequency and time resolution because of the Heisenberg uncertainty principle, the duration of the time window can be tailored to the analysis frequency. Thus, a spectrogram would consist of time slices depending, not only on the time coordinate, but also the frequency coordinate. In the particular case of a wavelet transform, one obtains an octave-band decomposition. However, there are more general orthonormal bases that allow a more irregular tiling of the time-frequency plane, with a lower bound set on the area of a tile by the Heisenberg uncertainty principle.

It is proposed here that a multiresolution form of decomposition be used for articulatory synthesis, as well as analysis. This would involve decomposing the time-dependent part of the equations of motion for air in the vocal tract into a general orthonormal basis. The vocal tract could still be divided into small tube sections for the spatial discretization, if desired. The matrices describing the transformation of pressure and volume velocities from one section to another (Sondhi and Schroeter's chain matrices) would be written in new orthonormal coordinates. In one possible implementation of a multiresolution synthesis area functions would be sampled at a fast rate, and this area function averaged over different intervals depending on the frequency scale of interest, with the higher frequencies requiring less duration for quasisteady conditions than the low frequencies. The time derivatives, including fractional derivatives, would be written in terms of the chosen orthonormal basis. While Fourier analysis transforms the derivative operator to a diagonal operator, the wavelet decomposition transforms the derivative into a sparse matrix. This sparse matrix can be used for fast computation of derivatives in the wavelet basis (Beylkin, 1993). Further, any noise sources can be can be shaped by bandpass filters composed of wavelets.

FOUR PARAMETER VOICE SOURCE

Articulatory synthesizers can have voice sources that are not controlled using articulatory parameters (e.g. Rubin, et al. 1981). Or they have voice sources that are continuum mechanical models of the self-oscillating folds and air flow in the laryngeal region (Ishizaka & Flanagan, 1972). The latter simulations can require too much computation time or produce poor voice quality in running speech. While the former voice sources can produce natural sounding voice, they are not controlled by articulatory parameters.

The cover-body model of the vocal folds is the starting place for a four parameter model of the voice source (Hirano, 1974). It is supposed that all aspects of voice quality can be determined by "...the relationship between the body and cover of the vocal fold" (Hirano, 1974, p 91). The cover-body picture of phonation has been expanded by others, most notably Titze (1994), who has constructed muscle activation plots (MAPs) for fundamental frequency control. (While these MAPs have largely been based on canine data, Titze's group has recently measured stress-strain relations for the human vocal ligament [Titze et al., 1994].) In these plots isofrequency contours are plotted against cricothyroid (CT) muscle activation and thyroarytenoid (TA) activation. However, for purposes of controlling the properties of the of the cover and body of the folds, the CT activation could be thought to represent any factor, intrinsic or extrinsic that controls the length of the cover and body, and stiffening both structures when they are lengthened. This change could be due to factors such as raising and lowering the larynx so that the larynx changes position along the spine, thus rotating the thyroid and cricoid relative to one another (Honda, 1995). Also, muscles who's primary effect is thought to be abductory and adductory motion of the folds can have an effect the length of the cover-body in ways analogous to the CT. On the other hand, TA activity, reduces the length of the cover-body complex, but by stiffening the body and relaxing the cover. There are many ways of attaining the same fundamental frequency using different combinations of CT and TA activation. However, these different combinations can often, if not always, be distinguished in other acoustic dimensions, including source amplitude and spectral content.

There are other ways to control fundamental frequency, source amplitude and spectral content. These include the degree of adduction and the transglottal pressure. The latter parameter is partly determined by what is happening in the upper vocal tract independent of the larynx. A tight constriction in the upper vocal tract and an open glottis will mean that transglottal pressure decreases. Thus, the four parameters in the proposed model of voice source control are the transglottal pressure, degree of abduction, (generalized) CT activity, and TA activity. These parameters should provide enough degrees of freedom to produce just about any voice quality. This would not be true if one of these parameters were omitted, and so this set could be considered minimal. Also, while it is not a detailed anatomical model of the larynx and its vibratory modes, it is sufficiently articulatory given the state of the art in articulatory synthesis.

ACKNOWLEDGMENTS

This work was supported by grant NIH grant DC-01247 to Haskins Laboratories. The description of task dynamics has benefited by discussions the author has had with Elliot Saltzman. Thanks to Phil Rubin and Doug Whalen for reviewing this work.

REFERENCES

Beylkin, G. (1993). Wavelets and fast numerical algorithms. In I. Daubechies (ed.) *Different Perspectives on Wavelets, Proceedings of Symposia in Applied Mathematics, Volume 47.* Providence: American Mathematical Society. 89-117.

Bickley, C., Stevens, K.N., and Williams, D.R. (1994). A framework for synthesis of segments based on articulatory parameters. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis.*

Browman, C.P. and Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *J. Phonetics, 20.* 27-38.

Coifman, R. (1991). Adapted multiresolution analysis, computation, signal processing, and operator theory. In *Proceedings of the International Congress of Mathematicians, Kyoto, 1990.* Tokyo: Springer-Verlag.

Coifman, R.R. and Wickerhauser, M.V. (1993). Wavelets and adapted waveform analysis. A toolkit for signal processing and numerical analysis. In I. Daubechies (ed.) *Different Perspectives on Wavelets, Proceedings of Symposia in Applied Mathematics, Volume 47.* Providence: American Mathematical Society. 119-153.

Davies, P.O.A.L., McGowan, R.S., and Shadle, C.H. (1993). Practical Flow Duct Acoustics. In I.R. Titze (ed.) *Vocal Fold Physiology: Frontiers in Basic Science.* San Diego: Singular Publishing Group, Inc.

Guenther, F.H. (1994). *Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production,* (Technical Report CAS/CNS-94-012). Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems, Boston, MA.

Hirano, M. (1974). Morphological structure of the vocal cord as a vibrator and its variations. *Folia Phoniatrica, 26.* 89-94.

Honda, K. (1995). Laryngeal and extra-laryngeal mechanisms of F0 control. In *Producing speech: Contemporary issues,* F. Bell-Berti and L.J. Raphael (Eds.). Woodbury, NY: AIP Press. pp. 215-232.

Ishizaka, K. and Flanagan, J.L. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *The Bell System Technical Journal, 51.* 1233-1268.

Lin, Q. (1994). Vocal-tract computation: How to make it robust and fast. *J. Acoust. Soc. Am., 96.* 2576-2579.

Maeda, S. (1982). A digital simulation method of the vocal-tract system. *Speech Communication, 1.* 199-229.

Mattingly, I.G. (1990). The global character of phonetic gestures. *J. Phonetics, 18.* 445-452.

McGowan, R.S. and Saltzman, E.L. (in press). Incorporating aerodynamic and laryngeal components into task dynamics. *J. Phonetics.*

Meyer, Y. (1993) *Wavelets, Algorithms & Applications.* Philadelphia, PA: Society for Industrial and Applied Mathematics

Parthasarthy, S. and Coker, C.H. (1992). On automatic estimation of articulatory parameters in a text-to-speech system. *Computer Speech and Language, 6.* 37-75.

Rose, R.C., Schroeter, J., and Sondhi, M.M. (1994). An investigation of the potential role of speech production models in automatic speech recognition. In *Proceedings of the International Conference on Spoken Language Processing, September 18-22, Yokohama, Japan., Volume 2.* pp. 575-578.

Rubin, P., Baer, T, and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *J. Acoust. Soc. Am., 93.* 1109-1121.

Saltzman, E.L. and Munhall, K.G. (1989). A dynamic approach to gestural patterning in speech production. *Ecological Psychology, 14.* 333-382.

Schroeter, J. and Sondhi, M.M. (1992). Speech coding based on physiological models of speech production. In S. Furui and M.M. Sondhi (eds.), *Advances in Speech Signal Processing.* New York: Marcel Dekker. pp. 231-268.

Shirai, K. and Kobayashi, T. (1986). Estimating articulatory motion from speech wave. *Speech Communication, 5.* 379-385.

Sondhi, M.M. and Schroeter, J. (1987). A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 35.* 955-967.

Titze, I.R. (1994). *Principles of Voice Production.* Englewood Cliffs: Prentice-Hall, Inc.

Titze, I.R., Min, Y.B. & Alipour-Haghighi, F. (1994). Stress-strain response of the human vocal ligament and its effect on F0 control. *J. Acoust. Soc. Am., 96.* 3324.

Wickerhauser, M.V. (1993). Best-adapted wavelet packet bases. In I. Daubechies (ed.) *Different Perspectives on Wavelets, Proceedings of Symposia in Applied Mathematics, Volume 47.* Providence: American Mathematical Society. 155-171.