SPEECH PRODUCTION AND PERCEPTION MODELS AND THEIR APPLICATIONS TO SYNTHESIS, RECOGNITION, AND CODING

A. Alwan, S. Narayanan, B. Strope, and A. Shen,

Department of Electrical Engineering, UCLA
405 Hilgard Ave.
Los Angeles, CA 90024

ABSTRACT

The last few decades have witnessed tremendous progress in the performance, reliability, and wide-spread use of speech-processing devices. Using mathematical models of human speech production and perception has been an important factor in the improved performance of these devices. In this tutorial paper, we review recent advances in speech production and perception modeling and summarize the challenges that lie ahead in developing fully parametric models.

1. INTRODUCTION

Quantitative models of human speech production and perception provide important insights into our speech production and perception mechanisms and lead to high-quality computer synthesis of speech, robust automatic speech recognition (ASR), and efficient speech and audio coders. These issues are of importance in the development of effective human-computer communications through the medium of human language.

Simplified linear models of speech production form the basis of several speech synthesizers [Klatt, 1987] and the most widely-used speech coder today: Code-excited-linear-prediction (CELP) [Schroeder and Atal, 1985]. Simple auditory models have been used successfully in optimizing the performance of speech and audio coders [Schroeder et al., 1979; Shen et al., 1995; Johnston, 1988]; these models are also embedded in the MPEG audio-coding standards [Brandenburg and Stoll, 1992]. Using auditory models as preprocessors has resulted in improving the performance of recognition systems in noise [Ghitza, 1986].

Supported in part by NSF, NIH, and ARPA/CSTO. Shrikanth Narayanan is now with ATT Bell Labs, and Albert Shen, with Intel. Inc., OR.

Recently, with rapid advances in hearing research, medical imaging, non-linear analysis, and computer technology, more sophisticated speech production and perception models are being proposed.

2. SPEECH PRODUCTION MODELS

Speech production research is directed towards better understanding and quantitative characterization of the acoustics, articulatory dynamics, and cognition of both normal and pathological human speech. Speech production modeling efforts, however, face at least two major challenges: (1) the lack of appropriate physical and physiological data which are crucial for developing quantitative models, and (2) the presence of articulatory and acoustic variabilities, both within and across speakers.

A great deal of attention is, hence, given to the development and use of novel measurement and instrumentation techniques that provide better insights into speech production mechanisms. Examples of such techniques include the use of magnetic resonance imaging (MRI) to study the 3D vocal-tract geometry, use of ultrasound [Stone, 1990], x-ray microbeam and electromagnetic midsagittal articulometer (EMMA) [Perkell, 1992] to study tongue dynamics, measuring linguopalatal interaction using electropalatography (EPG) [Hardcastle et al, 1989], and the use of microsensor-based aerodynamic and acoustic measurements in the vocal tract. Since most of these techniques provide information about certain specific attributes of speech production, data are typically collected using several of these techniques in parallel, if not simultaneously. Novel signal processing techniques are often required to process, visualize, and quantify the various physical and physiological speech production data thus obtained.

A major focus of speech production research is in modeling articulatory-acoustic relationships of speech sounds. Physically and physiologically-based models for speech acoustics are particularly important for developing high-quality speech synthesis and low bit rate (articulatory) coding. Significant progress has been recently made towards developing improved articulatory-acoustic models. Extensive data from human subjects using techniques such as MRI were used to obtain accurate quantitative and qualitative characterizations of the vocal-tract geometry for vowels and consonants [Baer et al., 1991; Narayanan et al., 1995]. The information obtained from these experiments, such as accurate vocal-tract dimensions, were then employed in modeling the acoustics of these sounds. For example, physically-motivated 1D acoustic models for fricatives, showing good agreement with the spectra of natural fricatives, were recently reported [Narayanan, 1995].

Numerical simulation of the acoustic wave propagation in actual 3D vocal-tract models, particularly for 'static' scenarios corresponding to sustained sounds, using techniques such as finite element and/or finite time-difference methods has received wide attention [Miki et al., 1994; Cummings et al., 1995]. This is largely due to the vast improvements in computational capabilities and the availability of more realistic vocal-tract data. An alternative to numerical simulations for studying flow problems in the 3D vocal tract is to use mechanical analogs. It should, however, be noted that simulation of speech *dynamics* in 3D models is an extremely challenging problem, which is as yet largely untackled.

Considerable interest has also been devoted to the inverse problem in speech production i.e., the identification of articulatory parameters from the acoustic speech signal. Solution to the inverse problem has potential applications in low bit rate coding and speech recognition [Schroeter and Sondhi, 1992; Deng and Sun, 1994]. It is, however, important to bear in mind that the unconstrained inverse problem is non-unique. That is, the acoustic to articulatory transformation is not one-to-one. Identifying appropriate constraints to solve uniqueness is not a solved problem. Various techniques such as neural networks [Shirai, 1993] and genetic algorithms [McGowan, 1994] have been proposed for solving the inverse problem although their viability is yet to be fully demonstrated.

Novel approaches such as those drawn from nonlinear dynamical systems (chaos) theory have recently found applications in speech analysis. Such analyses have provided further insights into the production of pathological speech [Titze et al., 1993] and consonant sounds such as fricatives which are characterized by turbulence formation in the vocal tract [Narayanan and Alwan, 1995].

Recently, there has also been an interest in understanding higher-level processing during speech production by monitoring brain activity using functional MRI [Shaywitz et al., 1995] in conjunction with other methods such as PET.

Considerable efforts directed towards modeling speech production are hoped to provide a clearer picture of the underlying mechanisms and help in the development of better engineering applications. The significance of efficient analysis schemes (identification of the articulatory parameters from the speech waveform) and synthesis schemes (generation of the acoustic speech waveform from the articulatory parameters) extends well beyond the realm of speech production modeling with clinical and linguistic implications.

3. SPEECH PERCEPTION MODELS

Much of the evidence for current auditory models are based on acoustic masking experiments. Static (or simultaneous) masking experiments have led to the definition of the critical bandwidth, and a non-uniform filter bank model of auditory perception [Fletcher, 1940; Zwicker et al, 1957; Zwicker and Terhardt, 1980]. Another well-quantified aspect of auditory processing is "loudness equalization" [Robinson and Dadson, 1956]. For example, the perceived loudness of two tones at different frequencies might be different even though their intensity levels (in dB SPL) are the same.

Dynamic (or non-simultaneous, forward) masking experiments [Plomb, 1964; Jesteadt et al, 1982; Moore and Glasberg, 1983] also provide significant insight for auditory modeling efforts. Specifically, forward masking experiments suggest that the relative levels of short-time spectral estimates are perceptually significant.

Further, perceptual experiments with spectrally complex speech-like stimuli indicate that human audition is particularly sensitive to the frequency location of local spectral peaks [Klatt, 1982]. Measurements of the timing detail of individual inner hair cell responses reveal firing responses which track dominant spectral features [Delgutte and Kiang, 1984] (instead of exclusively representing the spectral content at the fiber's center or best frequency as a filter bank model might predict) suggesting a possible mechanism for this spectral peak sensitivity.

Our understanding of auditory processing of speech signals is at an early stage; one of the major challenges in the next few decades is unraveling and modeling the neural mechanisms involved in speech processing [Pickles, 1994].

Auditory Models and Automatic Speech Recognition

Early efforts to build automatic speech recognition machines occurred in the 1950s at Bell Laboratories [Davis et al, 1952]. These systems relied on identifying the spectral resonances of vowels within isolated digits. Despite tremendous research focus, and many significant technical advances, a robust speech recognition system that approaches human performance still does not exist. Research, therefore, continues.

Typical speech recognition systems involve two fundamental steps: short-term spectral analysis, followed by pattern comparison with representative templates (or statistical models of templates). In the 1970s, LPC-based spectral analysis [Itakura, 1975], and the application of dynamic programming techniques to pattern comparison [Vintsyuk, 1968] led to successful isolated word recognition systems. Today many systems use Mel-Frequency Cepstral Coefficients (and their temporal derivatives) for spectral analysis, and Hidden Markov Models (HMM) of the templates for pattern comparison [Rabiner, 1989]. Mel-Frequency Cepstral Coefficients (MFCC) are defined as the DCT of log spectral estimation obtained with a critical bandwidth-like non-uniform filter bank model [Davis and Mermelstein, 1980]. The DCT provides an orthogonal transformation to a vector space with better energy compaction, which therefore requires fewer coefficients per acoustic vector. Including temporal derivatives of the cepstral coefficients in the acoustic observation vector is to account for dynamic sensitivity [Rabiner and Juang, 1993].

Providing a rigorous stochastic framework, HMMs and the techniques to train and apply them, have led to successful large-vocabulary speaker-independent systems [eg., Lee et al, 1991]. Unfortunately the performance of these systems still degrades significantly when the acoustic environment (amount and type of background noise, reverberation, competing sources, etc.) differs from the training data.

HMM systems [Rabiner, 1989] rely on a fundamental stochastic identity (Bayes Theorem). HMMs provide the probability of an acoustic observation for each specific template model. We then choose the template model which generates the highest probability of the acoustic observation occurring.

If we are trying to answer the question: "What was most likely said?" it seems reasonable that we also want to characterize statistically the acoustic distinctions between sounds. Although recent efforts are addressing this issue, the formal derivation of a computationally tractable training procedure is much more difficult, and remains an open question.

Using MFCC and their temporal derivatives to represent the acoustic observation, current recognizers incorporate a suitable model of frequency selectivity and a rough approximation of short-term auditory adaptation. A number of researchers have proposed models with more explicit short-term auditory adaptation [Seneff, 1988; Lyon and Mead 1988; Hermansky and Morgan 1994], however parameterizing these models so that they are quantitatively consistent with measurable top-level functionality remains a consistent challenge.

The recent system we developed includes a model of short-term auditory adaptation parameterized through a series of perceptual forward-masking experiments, and a novel cepstral processing technique to isolate local spectral peaks [Strope, 1995]. Figure 1 illustrates different representations for the digits one, three, and nine, spoken by a male speaker. The top part of the figure is the time waveform, the middle part, a spectrogram computed using cepstral coefficients derived from a linear prediction model, and the third is a "perceptual" spectrogram computed using our dynamic perceptual model. Note how the perceptual representation highlights onsets and spectral transitions which are perceptually important. Using a dynamic-programming based recognition system, this representation leads to significantly more robust recognition performance when compared to common cepstral and cepstral derivative representations. Currently, we are also evaluating this representation with an HMM system.

Stochastic modeling provides a powerful tool to train recognition systems based on a sequence of acoustic observations. Future efforts will undoubtedly continue to improve these stochastic models and training techniques. We should also expect, however, increased performance by incorporating acoustic observation sequences that are more consistent with human perception.

Auditory Models and Speech and Audio Coding

Speech coding is the process of obtaining a compact representation of voice signals for efficient transmission and storage. Obtaining an efficient method of transmitting speech signals over band-limited wired and wireless channels drives the research in speech coding. Today, in a software-controlled digital format, speech coders have become an essential component in telecommunications and in the multimedia infrastructure [Jayant and Cox, 1993]. Commercial systems which rely heavily on efficient speech coding include cellular communication, video conferencing, digital simultaneous voice and data (DSVD), as well as numerous PC-based games and multimedia applications.

To measure the overall effectiveness of digital speech coders, three metrics are commonly used: complexity, perceptual quality, and bit rate. The complexity of a speech coding algorithm is reflected in the number of machine instructions executed when operating in a real time environment, measured in millions of instructions per second (MIPS). The second criterion is perceptual quality. A common metric used to measure perceptual quality is the Mean Opinion Score (MOS). This measure is an average score derived from subjective listening tests using a five-point scale. A score of one is assigned to speech with 'poor' quality, while a score of five is given to speech with 'excellent' quality. The third metric is the operating bit rate of a coder, which is the rate at which data must be transmitted or stored in order to effectively reconstruct the original speech signal [Jayant and Noll, 1984].

Auditory modeling can aid in the design of speech and audio coders in two important ways: 1) developing an objective measure to rapidly and reliably assess the performance of a coder instead of conducting time-consuming and expensive subjective listening tests [Wang et al., 1992], and 2) optimizing the performance of perceptually-based speech and audio coding schemes [Shen, 1994; Shen et al, 1995].

Simplified auditory models have been used successfully in both areas (developing objective measures and designing speech and audio coders). These simplified models view speech as a sequence of unrelated static segments and exploit, predominantly, static masking effects. A more complete auditory model, especially one which takes into account *dynamic* spectral distortions, will undoubtedly have an important impact in speech coding research.

4. SUMMARY AND CONCLUSION

It is clear that our understanding of speech production and perception mechanisms has improved tremendously in the past few decades. Further multidisciplinary research in signal processing, acoustics, psychoacoustics, linguistics, imaging, and auditory physiology are needed to better model these mechanisms.

5. REFERENCES

[1] Baer, T., Gore, J.C., Gracco, L.C., and P.W. Nye, "Analysis of vocal-tract shape and dimensions using MRI: Vowels", JASA 90, 799-828, 1991.

[2] Brandenburg, K. and G. Stoll. "The ISO/MPEG Audio Codec: A Generic Standard for Coding of High Quality Digital Audio." Audio Engineering Society Preprint 3336, 1992.

[3] Cummings, K., Maloney, J., and M. Clements, "Modeling Speech Production using Yee's Finite Difference Method," ICASSP Proc., Vol. 1, pp. 672-675, 1995.

[4] Deng, L. and D.X. Sun, " A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," JASA 95, 2702-2719, 1994.

[5] Davis, K. H., Bidduplh, R., Balashek, S. "Automatic Recognition of Spoken Digits," JASA, 24, 637-642, 1952.

[6] Davis, S. B., Mermelstein P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. ASSP 28, August, 357-366, 1980.

[7] Delgutte, B. and Kiang, N. Y. S. "Speech coding in the auditory nerve: I. Vowel-like sounds," JASA 75, 866-878, 1984.

[8] Fletcher, H. "Auditory Patterns," Rev. Mod. Physics 12, 47-65, 1940.

[9] Ghitze, O. "Auditory nerve representation as a front-end for speech recognition in a noisy environment," Computer Speech and Language, 1(2):109-130, 1986.

[10] Hardcastle, W.J., Jones, W., Knight, C., Trudgeon, A. and G.Calder, "New developments in electropalatography: A state of the art report," Clinical Linguistics and Phonetics, vol.3, pp. 1–38, 1989.

[11] Hermansky, H., and N. Morgan "RASTA processing of speech," IEEE Trans. Speech and Audio Proc., vol. 2, no. 4, pp. 578-589, 1994.

[12] Itakura, F. "Minimum Prediction Residual Applied to Speech Recognition," IEEE Trans. ASSP 23, February, 67-72, 1975.

[13] Jayant, N.S. and P. Noll. *Digital Coding of Waveforms: Principles and Applications to Speech and Video.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[14] Jayant, N.S. and R.V. Cox. "Speech Processing." IEEE Communications Magazine. Vol. 31, No. 11, 1993.

[15] Jesteadt, W., Bacon, S., and Lehman, J. "Forward Masking as a function of frequency, masker level, and signal delay," JASA 71, 950-962, 1982.

[16] Johnston, James D. "Transform Coding of Audio Signals Using Perceptual Noise Criteria." IEEE JSAC, Vol. 6, No. 2, 1988.

[17] Kates, J., "An adaptive digital cochlear model," Proc. IEEE ICASSP, 3621-3624, Toronto, 1992.

[18] Klatt, D. "Prediction of perceived phonetic distance from critical-band spectra: a first step," Proc. ICASSP, Paris, 1278-1281, 1982.

[19] Klatt, D. "Review of text-to-speech conversion for English," JASA vol. 82, no. 3, pp. 737-793, 1987.
[20] Lee, K. F., Hon, H. W., and Huang, X. "Speech recognition using Hidden Markov Models: a CMU perspective," Speech Communication, 9, 497-508, 1991.
[21] Lyon, R. F., and Mead, C. "An Analog Electronic Cochlea," IEEE Trans. on Acoust., Speech, and Sig. Proc. 36, 1119-1133, 1988.
[22] McGowan, R.S., "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm," Speech Comm., 14, 19-48, 1994.
[23] Miki, N., Badin, P., Ngoc, P-T, and Y. Ogawa, "Vocal tract model and 3-dimensional effect of articulation", ICSLP Proc., Yokohoma, Japan, 167-170, 1994.
[24] Moore, B. C. J., and Glasberg, B. R."Growth of forward masking for sinusoidal and noise maskers as a function of signal delay; implications for suppression in noise," JASA 73, 1249-1259, 1983.
[25] Narayanan, S.S. "Fricative consonants: an articulatory, acoustic, and systems study," unpublished Ph.D. Dissertation, Dept. of Electrical Engineering, UCLA, June, 1995.
[26] Narayanan, S.S., and A.A. Alwan, "A nonlinear dynamical systems analysis of fricative consonants", JASA 97, 2511-2524, 1995.
[27] Narayanan, S.S., Alwan, A.A., and K. Haker, "An articulatory study of fricative consonants using MRI", to appear in the September issue of JASA, 1995.
[28] Perkell, J., "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements", JASA 6, 3078–3096, 1992.
[29] Pickles, J. *An Introduction to the Physiology of Hearing.* Second Edition, Academic Press, London, 1994.
[30] Plomb, R. "Rate of Decay of Auditory Sensation," JASA 36, 277-282, 1964.
[31] Rabiner, L. R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, 77, February, 257-286, 1989.
[32] Rabiner, L., and Juang, B. H. *Fundamentals of Speech Recognition.* Prentice-Hall, New Jersey, 1993.
[33] Robinson, D., and R. Dadson, "A redetermination of the equal-loudness relations for pure tone," Brit. J. Appl. Phys., pp 166-181, 1956.
[34] Schroeder, M. R., Atal, B. S., and Hall, J. L. "Optimizing digital speech coders by exploiting masking properties of the human ear," JASA, 66, 1647-1652, 1979.
[35] Schroeder, M.R. and B.S. Atal. "Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates." Proc. IEEE ICASSP, pp. 937-940, 1985.
[36] Schroeter, J., and M. M. Sondhi, "Speech Coding Based on Physiological Models of Speech Production", in *Advances in Speech Signal Processing*, edited by S. Furui and M. M. Sondhi, Marcel Decker, New York, 231-268, 1992.
[37] Seneff, S., "A joint synchrony/mean-rate model of auditory speech processing," J. Phonetics, vol. 16, no. 1m pp 55-76, 1988.
[38] Shaywitz, B.A. et al., "Sex differences in the functions organizational of the brain for language," Nature, 373, 607-609, 1995.
[39] Shen, A. "Perceptually-based subband coding of speech signals," unpublished Master's thesis, Dept. of Electrical Engineering, UCLA, June, 1994.
[40] Shen, B. Tang, A. Alwan, and G. Pottie, "A Robust and Variable-Rate Speech Coder," Proc. IEEE ICASSP 1995, Vol. I, 249-252.
[41] Shirai, K. "Estimation and generation of articulatory motion using neural networks", Speech Comm. 13, 45-51, 1993.
[42] Stone, M. "A three-dimensional model of tongue movement based on ultrasound and X-ray microbeam data", JASA 87, 2207-2217, 1990.
[43] Strope, B. "A Model of Dynamic Auditory Perception and its Application to Robust Speech Recognition," unpublished Master's thesis, Dept. of Electrical Engineering, UCLA, June, 1995.
[44] Titze, I. R., Baken, R., and H. Herzel, "Evidence of chaos in vocal fold vibration," in *Vocal fold physiology: New Frontiers in basic science,* 143-188, Singular, San Diego, 1993.
[45] Vintsyuk, T. K. "Speech Discrimination by Dynamic Programming," Kibernetika, 4, January-February, 81-88, 1968.
[46] Wang, S., Sekey, A., and A. Gersho, "An objective measure for predicting subjective quality," IEEE JSAC, Vol. 10, No. 5, 1992.
[47] Zwicker, E., Flottorp, G., and Stevens, S. "Critical Bandwidth in Loudness Summation," JASA 29, 548-557, 1957.
[48] Zwicker, E., Terhardt, E. "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," JASA 68, 1523-1525, 1980.