Computational Models for Auditory Speech Processing

Li Deng

Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
email: deng@crg6.uwaterloo.ca

Summary. Auditory processing of speech is an important stage in the closed-loop human speech communication system. A computational auditory model for temporal processing of speech is described with details of numerical solution and of the temporal information extraction method given. The model is used to process fluent speech utterances and is applied to phonetic classification using both clean and noisy speech materials. The need for integrating auditory speech processing and phonetic modeling components in machine speech recognizer design is discussed within a proposed computational framework of speech recognition motivated by the closed-loop speech chain model for integrated human speech production and perception behaviors.

1. Introduction

Auditory speech processing is an important component in the closed-loop speech chain underlying human speech communication. The roles of this component are to receive and to subsequently transform the raw speech signal, which is often severely distorted and significantly modified from that generated by the human speech production system, into suitable forms that can be effectively used by the linguistic decoder or "interpreter" based on its internal "generative" model for optimal decoding of the phonologically-coded messages. The computational approach to auditory speech processing to be described in this paper has been developed from a detailed biomechanical model of the peripheral auditory system up to the level of auditory nerve (AN) [5, 2, 7]. The processing stages in the auditory pathway beyond the AN level will not be covered here and interested readers are referred to a few recent, excellent review articles (e.g. [1, 9]) and to some preliminary work published in [8].

The component modeling approach to auditory speech processing described in this paper appears to be a rightfully viable one at the present stage of the auditory-model development. This contrasts the development of speech production models where global modeling has been the main focus [4]. Development of appropriate statistical structures in global auditory models in the future will rely on considerable further efforts in the development of component models.

2. A nonlinear computational model for basilar membrane wave motions

The computational model of the basilar membrane (BM) used for speech processing is of a nonlinear, transmission-line type, which has been motivated by a number of

key biophysical mechanisms known to be operative in actual ears [5, 2]. The final mathematical expression which succinctly summarizes the model is the following nonlinear partial differential equation (wave equation):

$$\frac{\partial^2}{\partial x^2}\left(m\frac{\partial^2 u}{\partial t^2}+r(x,u)\frac{\partial u}{\partial t}+s(x)u-K(x)\frac{\partial^2 u}{\partial x^2}\right)-\frac{2\rho\beta}{A}\frac{\partial^2 u}{\partial t^2}=0, \qquad (1)$$

where $u(x,t)$ is BM displacement function of time along longitudinal dimension x; $m, s(x)$, and $r(x,u)$ are model parameters for BM unit mass (constant), stiffness (space dependent), and damping (space and output dependent), respectively, and $K(x)$ is BM lateral stiffness coupling coefficient. Nonlinearity of the model comes from output-dependent damping parameter $r(x,u)$, whose biophysical mechanisms and functional significance in speech processing have been discussed in detail in [5, 2, 7]. Input speech waveforms or other arbitrary acoustic inputs to the model enter into the partial differential equation (1) via the boundary condition at $x = 0$ (stapes).

The derivation of the above model is based on 1) Newton's second law; 2) fluid mass conservation law; 3) mechanical mass-spring-damping properties of the basilar membrane; and 4) outer hair-cell motility properties (which produce nonlinear damping $r(x,u)$). The model's output, $u(x,t)$, can be viewed as nonlinear traveling waves along the longitudinal dimension of the BM, or as a highly-coupled bank of nonlinear filter outputs. Both the derivation and the wave properties of this BM model are very similar to those of the partial differential equation governing vocal tract acoustic wave propagation (except the latter typically gives linear wave propagation). [1]

3. Frequency-domain and time-domain computational solutions to the BM model

The nonlinear partial differential equation (1) does not have analytic solution for arbitrary acoustic input signals. The only viable approach to obtaining model outputs appears to be computational means by numerical solution. Two methods of numerical solution, frequency-domain and time-domain methods based on the finite-difference scheme , will be described with their respective strengths and weaknesses discussed.

The frequency-domain method is significantly faster than the time-domain counterpart, but requires batch processing (non real-time) and linearization of the BM model. Linearization of the BM model results in some degrees of loss in the model

[1] In this parallel, the mechanical property of the BM which consists of a damped mass-spring system causing BM vibration is analogous to the vocal tract wall vibration arising also from a damped mass-spring system . The same Newton's second law and mass conservation law lead to wave properties of the BM traveling wave and of the vocal tract acoustic wave.

solution's accuracy. This, however, can be somewhat but not fully mitigated by using adaptive linearization [2].

When Eqn.(1) is linearized by eliminating output-dependence of the damping term $r(x,u)$, frequency-domain solution of the model can be obtained using Fourier transforms:

$$\begin{aligned} u(x,t) &\longleftrightarrow u(x,j\omega), \\ \frac{\partial u(x,t)}{\partial t} &\longleftrightarrow jwu(x,j\omega), \\ \frac{\partial^2 u(x,t)}{\partial t^2} &\longleftrightarrow -\omega^2 u(x,j\omega). \end{aligned}$$

This turns Eqn. (1) into an ordinary differential equation:

$$\frac{d^2}{dx^2}\left\{\left(-m\omega^2 + s(x) + j\omega r(x)\right)u - k(x)\frac{d^2u}{dx^2}\right\} + \frac{2\rho\beta}{A}\omega^2 u = 0. \qquad (2)$$

Numerical solution of the above frequency-domain model by the finite-difference method requires that the spatial dimension be represented by a finite number of discrete points. The solution is obtained for the displacement of the BM, $u(x,j\omega)$, as a function of the distance from the stapes, x, for selected input frequencies, ω. To discretize the frequency-domain model, the derivatives in Eqn.(2) are approximated by the conventional central differences:

$$\frac{du}{dx} = \frac{u_{i+1} - u_{i-1}}{2\Delta x},$$

$$\frac{d^2u}{dx^2} = \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta x)^2},$$

$$\frac{d^4u}{dx^4} = \frac{u_{i+2} - 4u_{i+1} + 6u_i - 4u_{i-1} + u_{i-2}}{(\Delta x)^4}.$$

This then turns ordinary differential equation (2) into a linear algebraic equation, which can be solved by straightforward matrix inversion to give $u(x,j\omega)$. The time-domain output is finally obtained by taking inverse Fourier transform of $u(x,j\omega)$, one for each discrete point along the x dimension.

The time-domain numeric solution allows on-line processing, and solve arbitrarily complex nonlinear BM model without performing model linearization. But the computational load is significantly greater than the frequency-domain method since one matrix inversion is required for each sample of speech. The reason for the computational load is that we can no longer use Fourier transform due to nonlinear element(s) in the model. Hence, both time and space variables need to be discretized. After the discretization, we use the following finite difference approximation to all partial derivatives, from order one to order four, in Eqn. (1):

$$\frac{\partial u}{\partial t} = \frac{u_i^{n+1} - u_i^n}{\Delta t},$$

$$\frac{\partial^2 u}{\partial t^2} = \frac{u_i^{n+1} - 2u_i^n + u_i^{n-1}}{(\Delta t)^2},$$

$$\frac{\partial u}{\partial x} = \frac{u_{i+1}^n - u_i^n}{\Delta x},$$

$$\frac{\partial^2 u}{\partial x^2} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{(\Delta x)^2},$$

$$\frac{\partial^4 u}{\partial x^4} = \frac{u_{i+2}^n - 4u_{i+1}^n + 6u_i^n - 4u_{i-1}^n + u_{i-2}^n}{(\Delta x)^4},$$

$$\frac{\partial^3 u}{\partial t \partial x^2} = \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1} - u_{i+1}^n + 2u_i^n - u_{i-1}^n}{\Delta t \ \ (\Delta x)^2},$$

$$\frac{\partial^4 u}{\partial t^2 \partial x^2} = \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1} - 2u_{i+1}^n}{(\Delta t)^2 \ \ (\Delta x)^2} \frac{+4u_i^n - 2u_{i-1}^n + u_{i+1}^{n-1} - 2u_i^{n-1} + u_{i-1}^{n-1}}{(\Delta t)^2 \ \ (\Delta x)^2}.$$

This turns the partial differential equation into a large algebraic equation with the solution variable $u(x,t)$ indexed by both time t and space x. The numerical procedure proceeds by first fixing each time t index and finding the solution for u as a function of space index x via matrix inversion . Then, by advancing time one sample after another, the entire solution for $u(x,t)$ is obtained.

The above solution has been used to process a large amount of speech data (rf. [7, 8]). Theoretical work on stability analysis of the model solution, which is essential to guarantee successful use of the model for automatic processing large-sized data, has been carefully carried out in the work reported in [6].

4. Interval analysis of auditory model's outputs for temporal information extraction

The BM model's output obtained by the finite-difference method described in the preceding section is used as the input to the inner hair cell model, which consists of hyperbolic tangent compression followed by low-pass filtering. The final stage of the auditory model is for the AN-synapse, which receives the input as the inner hair cell model's output. The AN-synapse consists of pools of neurotransmitters, separated by membranes of varying permeability, which simulate the temporal adaptation phenomenon experimentally observed in the AN.

The above composite auditory model's output is an array of temporally varying AN firing probabilities in response to input speech sounds to the BM model. This output is subject to an interval analysis for temporal information extraction. The analysis is based on construction of the Inter-Peak-Interval Histogram (IPIH) of the

dominant intervals measured from autocorrelation of 10-ms segments of the auditory model's output. In the IPIH construction, increment of each bin in the histogram is multiplied by the amplitude of the peak at the start of the corresponding interval. [2] Further, in the IPIH construction, a fixed number of intervals in the autocorrelation function are counted which are common across all AN output channels. This gives rise to approximately exponential temporal analysis windows, with the low-frequency channels occupying longer windows than the high-frequency channels. Finally, to reduce the data rate, the IPIHs constructed for all AN output channels are amalgamated, resulting in a single histogram per time frame. [3] Figure 1 shows an example of the process in the IPIH construction described above.

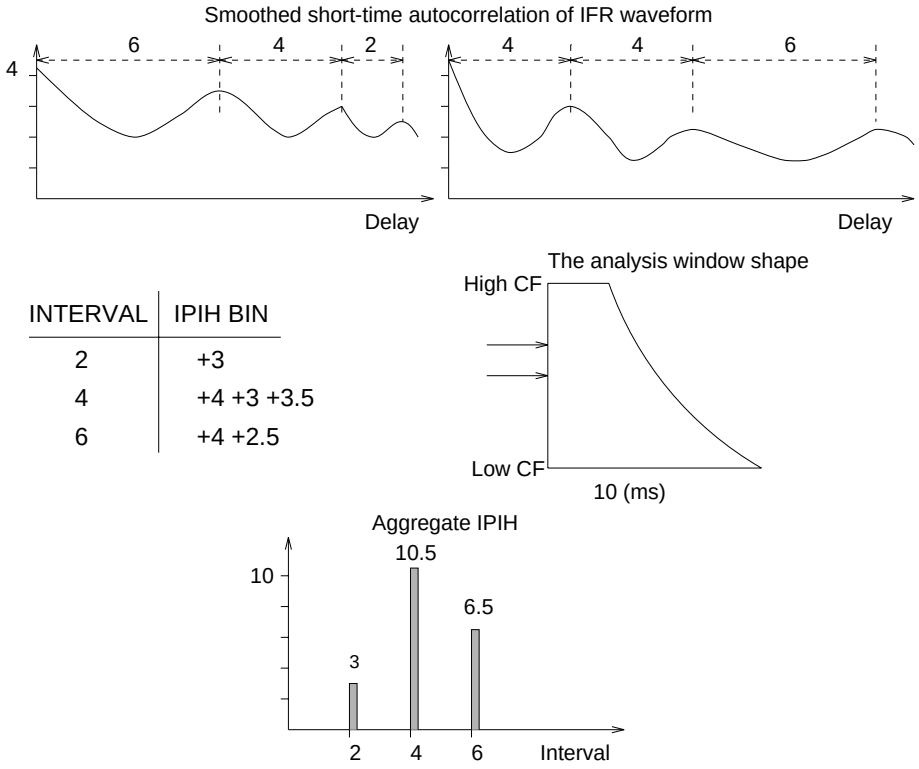

FIGURE 1. Construction of IPIH from the autocorrelation of the modeled AN instantaneous firing rate function

[2] This permits the IPIH to code the firing rate information in addition to the otherwise temporal information only.

[3] Note that the length of the time frame is frequency dependent (i.e. conditioned on the AN channel' center frequency).

5. IPIH representation of clean and noisy speech sounds

We have run the auditory model and carried out the consequent IPIH analysis on a number of utterances in the TIMIT database which cover a wide range of acoustic phonetic classes in American English. The model has been run for both clean speech and speech embedded in additive noise. A few examples are provided here to illustrate how various classes of speech sounds are represented in the form of IPIH constructed from the time-domain output of the auditory model as a temporal-nonplace code, and to show robustness of the representation to noise degradation.

Plotted in Figure 2 are the IPIHs for clean utterance *heels* (a) and *semi* (b), respectively, both presented to the auditory model at 69 dB SPL. The prominent acoustic characteristics of these utterances are the wide range of the formant transitions in the vocalic segments. For [iy] in *heels*, F2 moves drastically down from near 2100 Hz toward near 1300 Hz (F2 of the postvocalic [l]); this acoustic transition is reflected in the corresponding peak movement in the IPIH from about 0.48-ms interpeak interval (starting at 60 ms) to the interval of 0.75 ms (ending at around 200 ms). Similarly, the slow rising F1 transition in acoustics is represented as the slow falling IPIH peaks. For [ay] in *semi*, the rising F2 from about 1200 Hz to 2000 Hz is reflected in the falling IPIH peak from around 0.85-ms to 0.5-ms.

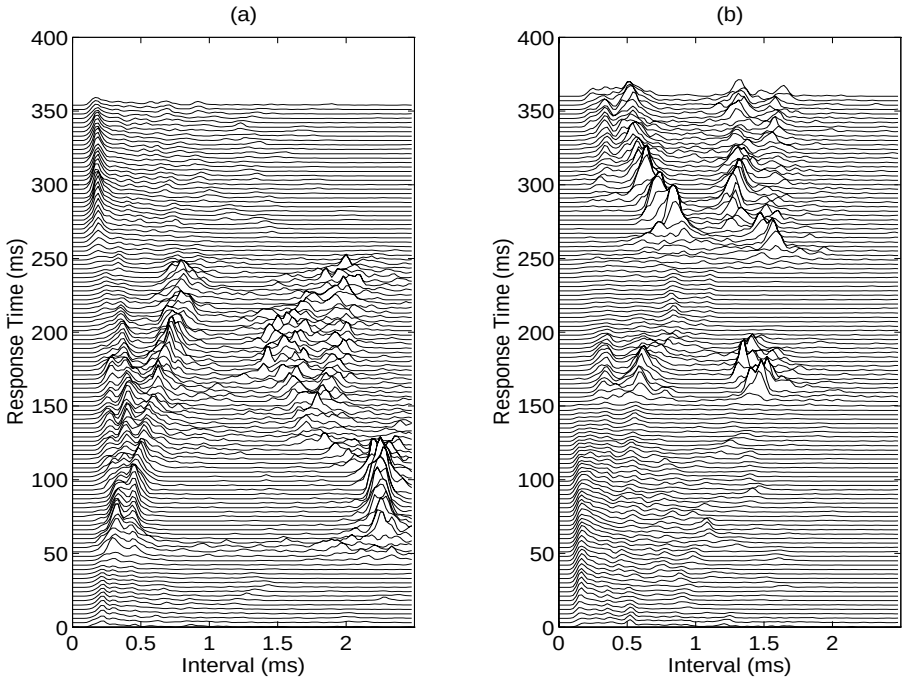

FIGURE 2. Modeled IPIHs for words (a) *heels* (b) *semi*

We have produced and analyzed the IPIHs for the words from several TIMIT sentences in much the same qualitative way as described above. From the analysis we find that all the significant acoustic properties of all classes of American English sounds that can be identified from spectrograms can also be identified, albeit to a varying degrees of modification, from the corresponding IPIH.

To evaluate noise robustness of the speech representation in terms of the interval statistics collected from the auditory-nerve population, we performed the identical IPIH analysis for the speech sounds identical to the ones described above except adding white Gaussian noise with 10-dB signal-to-noise-ratio (SNR) into the speech stimuli before running the auditory model. The resulting IPIHs for noisy versions of the utterances, *heels* and *semi* of Figure 2, are shown in Figure 3. A comparison between the IPIHs in Figures 2 and 3 shows that aside from some relatively minor distortions in the nasal murmur and in the aspiration, the major characteristics in the IPIH representation for the clean speech have been well preserved. In contrast to the above IPIH-based temporal representation in the auditory domain, the differences in the acoustic (spectral) domain between the clean and noisy versions of the speech utterances are found to be vast (not shown here).

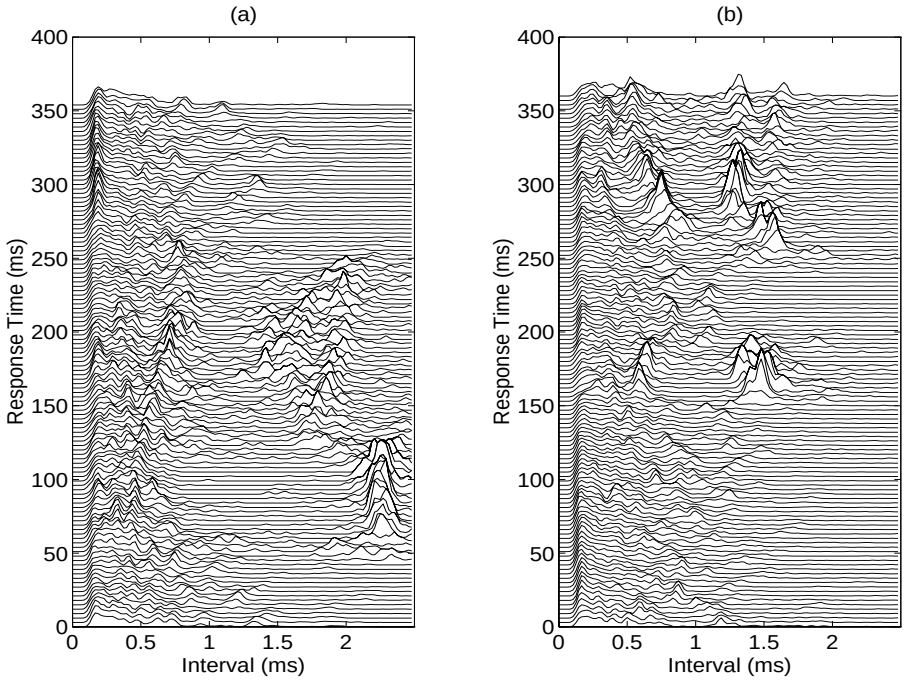

FIGURE 3. Modeled IPIH for words (a) *heels* (b) *semi* embedded in white noise with 10-dB SNR

.

6. Speech recognition experiments

The IPIH speech analysis results we have obtained demonstrated that the IPIH-based temporal representation preserves major acoustic properties of the speech utterances for all classes of English sounds in the magnitude-spectral domain, and that such a representation is robust to additive noise. One additional advantage of such a temporal representation over the conventional spectral representation in speech analysis is that the frequency resolution and time resolution can be controlled independently, rather than being constrained by an inverse, trade-off relationship. In our IPIH analysis, the time resolution is controlled by the frame size and by the overlap between adjacent frames, while the frequency resolution is independently determined by the number of cochlear channels set up in the model and by the bin width used to construct the IPIH. In principle, both the time and frequency resolutions can be increased simultaneously with no limits.

Despite these advantages, the IPIH-based temporal representation contains a much greater data dimensionality than that from the conventional magnitude-spectral analysis. Unfortunately, the current speech modeling methodology has not been advanced to the extent that the large data dimensionality required by the auditory temporal representation can be adequately accommodated and the data complexity associated with the large dimensionality be faithfully modeled. As such, heuristics-driven data dimensionality and complexity reduction methods have to be devised in order to interface the temporal representation of speech to any type of speech recognizer currently available.

Details of the experiments designed to evaluate the IPIH-based auditory representation are reported in [10]. The speech model embedded within the recognizer used in the experiments is the conventional, context-independent, stationary-state mixture HMM. This model requires that 1) the data inputs be organized to form a vector-valued sequence; 2) each vector in the sequence (i.e. a frame) contain an identical, relatively small number of components; and 3) the temporal variation of the vector-valued sequences be sufficiently smooth (except for occasional Markov state transitions which occur at a significantly lower rate than the frame rate but greater than the sample rate). To meet these requirements, we transform the IPIH representation of speech according to the following steps. First, the IPIH associated with each 10-ms time window is divided into a set of interval bands corresponding to the critical bands in the frequency domain. Each band contains a number of histogram bins, ranging from one for the high-frequency IPIH points to 15 for the low-frequency points. Second, the maximum histogram count within each interval band of the IPIH is kept while throwing out the remaining histogram counts. These maximum histogram counts, one from each interval band, preserve the overall IPIH profile while drastically reducing the data complexity. Third, this simplified IPIH is subject to further data complexity reduction via a standard cosine transform.

In the evaluation experiments,the speech data consist of eight vowels ([aa], [ae], [ah], [ao], [eh], [ey], [ih], [iy]) extracted from the speaker-independent TIMIT corpus. Tokens of the eight vowels (clean speech) from 40 male and female speakers (a total of 2000 vowel tokens) are used for training and those from disjoint 24 male and

female speakers (a total of 1200 vowel tokens) for testing. Both clean vowel tokens and their noisy version created by adding white Gaussian noise with varying levels of SNR are used as training and test tokens. The performance results, organized as the vowel classification rate as a function of the SNR level and of the two types of the speech preprocessor (IPI-based one with solid line vs. benchmark, MFCC-based one with dashed line), are shown in Figure 4. The results demonstrate that the auditory IPI-based preprocessor consistently outperforms the MFCC-based counterpart over a wide range of the SNR level (0 dB to over 15 dB). Only for near-clean vowels (20-dB SNR level), the two preprocessors become comparable in performance. [4]

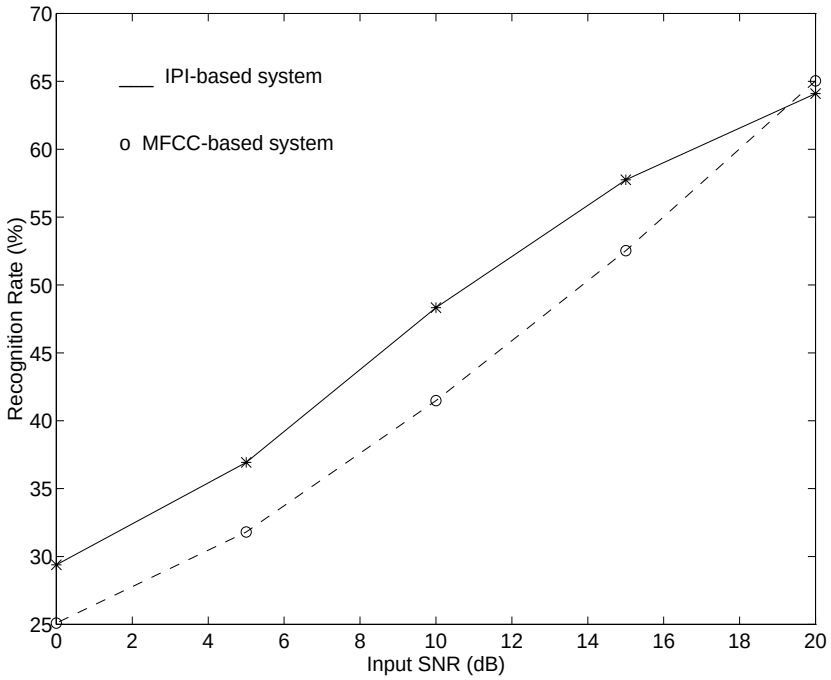

FIGURE 4. Comparative average classification rates for TIMIT vowels

7. Summary and discussions

With use of the computational auditory model described in this paper to process the speech utterances contained in the TIMIT database, it has been shown that not

[4] For evaluation experiments on other tasks and for details of the benchmark system, see [10].

only for limited and isolated speech tokens but also for a comprehensive range of manner classes of fluently spoken speech sounds, the auditory temporal representation on the basis of interval statistics collected from AN firing patterns preserves (with modification) the major acoustic properties of the speech utterances that can be identified from spectrograms. The temporal nature of the representation makes it robust to changes in the loudness level of the speech sounds and to the noise effect. The rate-level representation, which is closely related to the conventional spectral analysis, lacks such robustness.

Although the direction of exploring properties and constraints of the auditory system as a guiding principle for robust speech representation against noise effects in speech recognizer design appears to be promising, most experimental results (including ours and many other research groups' (too long to be listed here)) on recognition of noise-free speech have not been as successful as those for noisy speech compared with the conventional MFCC-based representation based more on traditional signal processing than on auditory properties. This is apparently caused by two competing factors working against each other. On the one hand, the independent specification of the time and frequency resolutions in speech preprocessing offered by the auditory interval-based representation allows potentially unlimited analysis resolutions for both time and frequency. On the other hand, however, the simultaneously greater resolutions enabled by the auditory representation are necessarily linked to a greater data dimensionality, causing problems for the speech modeling component of any current recognizer which requires relatively smoothed and redundancy-free patterns produced from the pre-processor. These two competing factors cannot be reconciled within the current HMM-based speech recognition framework. Any success in incorporating hearing science into speech recognition technology must come from *integrated* investigation of faithful auditory representation of speech and of the modeling component of the overall recognition system capable of taking full advantages of the information contained in the auditory representation. This integrated nature of the engineering system design can be closely paralleled with the biological counterpart of the closed-loop human speech communication system, where the auditorily received and transformed speech information must be fully compatible with what is expected from the listener's internal "generative" model approximating the speaker's linguistic behavior (and acting as an optimal decoder on the listener's part). Following this parallel, the integration of auditory representation and speech modeling components discussed here can be gratefully accomplished in the speech recognition architecture described in [3] which has been motivated by the global structure of the human closed-loop speech chain. Within this architecture, the role of computational auditory models will be to provide proper levels of auditory representation of the speech acoustics which will facilitate construction and learning of the nonlinear mapping between such representation and the internal production-affiliated variables. When this mapping is modeled within a global dynamic neural network system [4], then how to choose the output variables of the network to make model learning effective will place a

strongest demand on the level of details of auditory modeling which becomes a critical component of the integrated speech recognition architecture.

8. REFERENCES

[1] Delgutte B. (1997) "Auditory neural processing of speech," in *The Handbook of Phonetic Sciences*, W. J. Handcastle and J. Lavar (eds.), Blackwell, Cambridge, pp. 507-538.

[2] Deng L. (1992) "Processing of acoustic signals in a cochlear model incorporating laterally coupled suppressive elements," *Neural Networks*, Vol.5, No.1, pp.19–34.

[3] Deng L. (1998) "Articulatory features and associated production models in statistical speech recognition," this book.

[4] Deng L. (1998) "Computational models for speech production," this book.

[5] Deng L. and C.D. Geisler D. (1987) "A composite auditory model for processing speech sounds," *J. Acoust. Soc. Am.*, Vol. 82, No. 6, pp. 2001–2012.

[6] Deng L. and Kheirallah I. (1993) "Numerical property and efficient solution of a nonlinear transmission-line model for basilar-membrane wave motions," *Signal Processing*, Vol. 33, No. 3, pp. 269–286.

[7] Deng L. and Kheirallah I. (1993) "Dynamic formant tracking of noisy speech using temporal analysis on outputs from a nonlinear cochlear model," *IEEE Transactions on Biomedical Engineering*, Vol. 40, No. 5, pp. 456–467.

[8] Deng L and Sheikhzadeh H. (1996) "Temporal and rate aspects of speech encoding in the auditory system: Simulation results on TIMIT data using a layered neural network interfaced with a cochlear model," *Proc. European Speech Communication Association Tutorial and Research Workshop on the Auditory Basis of Speech Perception*, Keele Univ., U.K., pp. 75-78.

[9] Greenberg S. (1995) "Auditory processing of speech," in *Principles of Experimental Phonetics*, Ed. N. Lass, Mosby: London, pp. 362-407.

[10] Sheikhzadeh H. and Deng L. (1997) "Speech analysis and recognition using interval statistics generated from a composite auditory model," *IEEE Trans. Speech Audio Processing*, to appear.