THE AUDITORY PROCESSING AND RECOGNITION OF SPEECH

William Byrne, John Robinson, Shihab Shamma
Department of Electrical Engineering
University of Maryland College Park, MD 20742

ABSTRACT

We are carrying out investigations into the application of biophysical and computational models to speech processing. Described here are studies of the robustness of a speech representation using a biophysical model of the cochlea; experimental results on the representation of speech and complex sounds in the mammalian auditory cortex; and descriptions of computational sequential processing networks capable of recognizing sequences of phonemes.

INTRODUCTION

Systems for the automatic recognition of speech have in recent years derived many ideas and strategies from observations of the structure and processing modes of the nervous system, and specifically of the mammalian auditory system. Examples range from the adoption of cochlear-like processing as front-end analysis stages, to the use of artificial neural networks as adaptive pattern recognizers. For the last five years, we have been studying the functions and algorithms that facilitate the remarkable abilities of the auditory system to analyze, recognize, and localize complex sounds such as speech, music, and other environmental sounds. We have developed and used biophysical and computational models of the peripheral cochlear stages, of the intermediate central neural networks that extract various feature representations of the acoustic stimulus (e.g., as in speech phonemes), and of networks for the recognition of temporally ordered sequences (such as words and sentences). These models have been described in detail in [1,2,3,4,5]. Here, we shall outline a few of our recent investigations and the results that we have obtained.

NOISE ROBUSTNESS

Cochlear models in various forms are now commonly used in speech recognition systems. In many cases, they are severely simplified to reduce computational complexity, preserving only salient features of the original models, e.g., the psuedo-logarithmic frequency axis, critical-band filters, and the fast and/or slow adaptation (as AGCs). These and other processing steps have been justified in many elaborate and detailed experiments. One of the most desired features of cochlear processing has been *robustness to noise*, specifically, their supposed ability to provide a stable representation of the speech signal over a wide range of signal-to-noise ratios. Results from a few studies have so far been equivocal for many reasons, primary among them is the complexity of the systems tested which precluded clear separation of the causes of improvements and degradations. We have compared the noise immunity of cochlear representations to that of linear predictor coefficients (LPC), LPC cepstral coefficients, and discrete time Fourier Transform (DFT) spectra. Specifically, three investigations are performed: first, the distortion of each representation due to additive white noise is measured; in the second experiment, the robustness is measured through the deterioration in the vector quantizer performance of each representation; and finally, in the third experiment we measure the ability of each representation to discriminate speech sounds in noise.

Ninety sentences spoken by ten male speakers are taken from the phonetically labeled Icecream database and transformed into each of the four representations after upsampling from 16KHz to 20KHz (the cochlear model requires a 20KHz sampling rate). The cochlear model followed by two stages of lateral inhibition

[1] produces vectors of 128 tonotopically ordered elements from the 100Hz to 10KHz region of the basilar membrane. For all representations, a 20ms frame and 8ms step size are used, and, except for the cochlear model, a preemphasis of 1.0 and a Hamming window are applied.

The Log Area Ratios are obtained from the LPC coefficients of an order 28 predictor found via the autocorrelation method. The Log Area Ratios are used because of their appropriateness for vector quantization and mean square distortion measurements. The LPC cepstral coefficients are also computed via autocorrelation and the quefrency ranged from 0.0625 ms to 3.0625 ms. The spectrum is computed by a 256 point FFT with zero padding.

In the tests performed, noisy speech is obtained by adding white gaussian noise of the appropriate amplitude to the clean speech. The slight oversampling of the speech is taken into account in determining the noise amplitude. The signal to noise levels investigated are 24dB - 0dB in steps of 3dB.

FEATURE DISTORTION

The actual effect of additive noise on the various representations is measured first as

$$D_{Percent\ Distortion} = \frac{1/N \sum_{j=1}^{N} \|K(F(s_j)) - K(F(s_j + n_j))\|_2}{1/N \sum_{j=1}^{N} \|K(F(s_j))\|_2}$$

where $F(s_j)$ is the representation of frame j of the clean speech, N is the number of frames, and $s_j + n_j$ refers to a frame of speech with additive noise.

The Karhunen-Loeve transform, K, is computed for each representation from the autocovariance of the clean speech features. It is chosen as a means of reducing the dimension of the cochlear model in an optimum fashion. Since it also can be used to restrict measured data to a known signal space, it is applied to all representations so as not to give the cochlear model an unfair advantage. The eigenvectors corresponding to the 48 largest eigenvalues of the autocovariance matrix are chosen to form the transform kernel for both the spectral and cochlear representations. For the LPC and LPC cepstrum, all eigenvectors are retained. Note that if all eigenvectors are retained $\|K(F(s)) - K(F(s+n))\|_2 = \|F(s) - F(s+n)\|_2$ so the transform does not affect the distortion computation for the LPC and LPC cepstrum, and in practice, the spectral distortion is not reduced by the change in dimension.

The cochlear model suffers less distortion than the other representations at noise levels less than 9db, at which point it becomes parallel to the parametric models (Figure 1(a)).

VECTOR QUANTIZER DISTORTION

Another comparison among the different representations is through the effects of noise on the performance of vector quantizers (VQs) trained with clean speech. The effect of noise on the both VQ class distributions for each phoneme and the increase in codebook distortion are used as the measuring criteria.

Codebooks of 64 symbols are trained on clean speech and sample distributions of the VQ classes are formed for each phoneme at all noise levels. The similarity between the class distribution of the quantized clean speech, f_s, and the distribution of the quantized noisy speech, f_{s+n}, is measured by

$$D_{Distribution\ Distortion} = 1 - \frac{\sum_{i=1}^{64} f_s(i) \cdot f_{s+n}(i)}{\sqrt{\sum_{i=1}^{64} f_s^2(i) \sum_{i=1}^{64} f_{s+n}^2(i)}}$$

For presentation and comparison, the measurement of each representation is normalized by its 0db value. Only the results for the most frequently occurring vowel, /ey/, are given (Figure 1(b)), but the results for other phonemes, with the exception of stops, are essentially the same. In the case of the stops (and during silences), all representations seem to perform similarly.

Since the distribution of the VQ classes for particular phonemes is important to many statistical methods of speech recognition, the superior performance of the cochlear representation is significant. The class

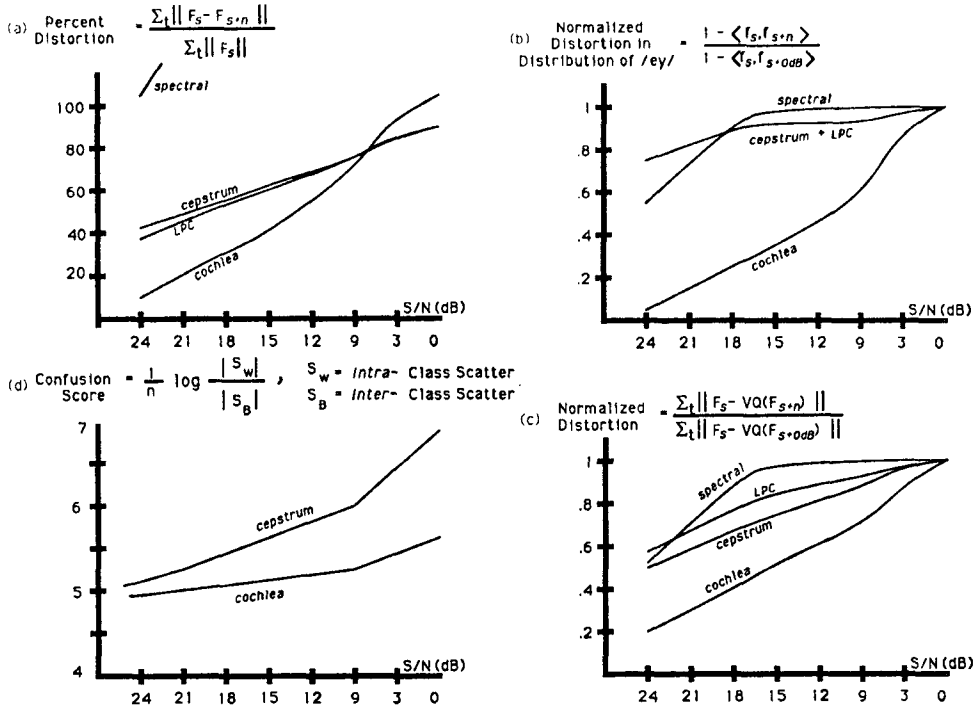

Figure 1: Distortion Due to Noise: (a) Percent Distortion (b) Change in VQ Class Distribution of /ey/ (c) Normalized VQ Distortion (d) Intra-Class vs Inter-Class Scatter

distributions show that the LPC and cepstrum, as noise increases, model all the speech sounds as noise, which the VQ labels as one of three or four classes. This happens also to the cochlear representation, but at higher noise levels.

An alternative way of measuring VQ performance is through the codebook distortion, defined as

$$D_{VQ\ Distortion} = 1/N \sum_{j=1}^{N} \|F(s_j) - VQ(F(s_j + n_j))\|_2$$

This is also computed for each phoneme, but only the composite results are presented here, normalized by the 0db distortion (Figure 1(c)).

A similar measure based on $1/N \sum_{j=1}^{N} \|VQ(F(s_j)) - VQ(F(s_j + n_j))\|_2$ is also computed. The results closely resemble those in Figure 1(c), but include a common bias due to the codebook distortion. The measures $D_{Distribution\ Distortion}$ and $D_{VQ\ Distortion}$ show that the cochlear model performs well at noise levels below 9db.

DISCRIMINATION ABILITY

The ability of the LPC cepstrum and cochlear model to discriminate between different phonemes in the presence of additive noise is an important performance measure in speech recognition. The phonetic labels

in the database are used to compute a variant of the Fischer Discrimination to compare the intra-class scatter to the inter-class scatter at each noise level. This measure favors representations in which features assigned to any particular phoneme are tightly clustered and distant from features assigned to other phonemes. The evaluation is given by

$$D_{Confusion\ Score} = 1/n \log \frac{\det S_W}{\det S_B}$$

where S_W and S_B are the intra-class and inter-class scatter matrices, respectively

$$S_W = \sum_{i=1}^{c} \sum_{x \in \chi_i} (x - m_i)(x - m_i)^t$$

$$S_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^t$$

and c is the number of phonemes, χ_i is the collection of all representations, x, labeled as the i^{th} phoneme, n_i is the cardinality of χ_i, and m and m_i are found by averaging all features and averaging all the features in χ_i, respectively.

Both the cepstrum and the cochlear model have similar discrimination performance at low noise levels (Figure 1(d)), but the cochlear model retains its performance better as the additive noise level increases.

DISCUSSION

Why is the cochlear representation performance superior to other representations? There are probably two sources: the first is the compression by the hair cell models; the second is the spectral extraction strategy - the lateral inhibitory network (LIN) - applied to the cochlear model output. Compression produces a well know effect of enhancing a signal in a noisy background (see [3]). In the cochlear models it is possible to apply strong compression without loss of spectral detail because the spectral information is encoded in the phase locked responses. The LIN utilizes this phase locking to extract a robust spectral estimate that can tolerate extreme compression. Such compression is not feasible for spectrogram representations since it completely destroys the spectral peaks and valleys.

AUDITORY NEUROPHYSIOLOGY

In the central auditory system, we are investigating the nature of the representation of complex acoustic spectra in the auditory cortex [4]. Recordings of unit responses along the isofrequency contours of the ferret primary auditory cortex reveal systematic changes in the symmetry of their receptive fields. At the center, units with narrow and symmetric inhibitory sidebands predominate. These give way gradually to asymmetric inhibition, with high frequencies (relative to the best frequency of the units) becoming more effective caudally, and weaker rostrally. This organization gives rise to a new columner organization in the primary auditory cortex that seems to encode spectral slopes and the symmetry of spectral peaks, edges, and envelopes. These columns are analogous to the well known orientation columns of the visual system. The implication of these findings is that in the perception and recognition of complex sounds special attention must be given to the representation of spectral gradients. We have simulated the receptive fields obtained in neurophysiological experiments and are in the process of examining in detail the representation of natural and synthetic stationary speech tokens in the responses of the cortex (Figure 2).

WORD RECOGNITION

Finally, we have been developing models of networks that can be used for the recognition of temporally-ordered sequences (e.g., phoneme sequences in a word) [5]. These networks are biologically plausible in that they do not require delay-lines to memorize the word prior to recognition. Instead, they function in a manner

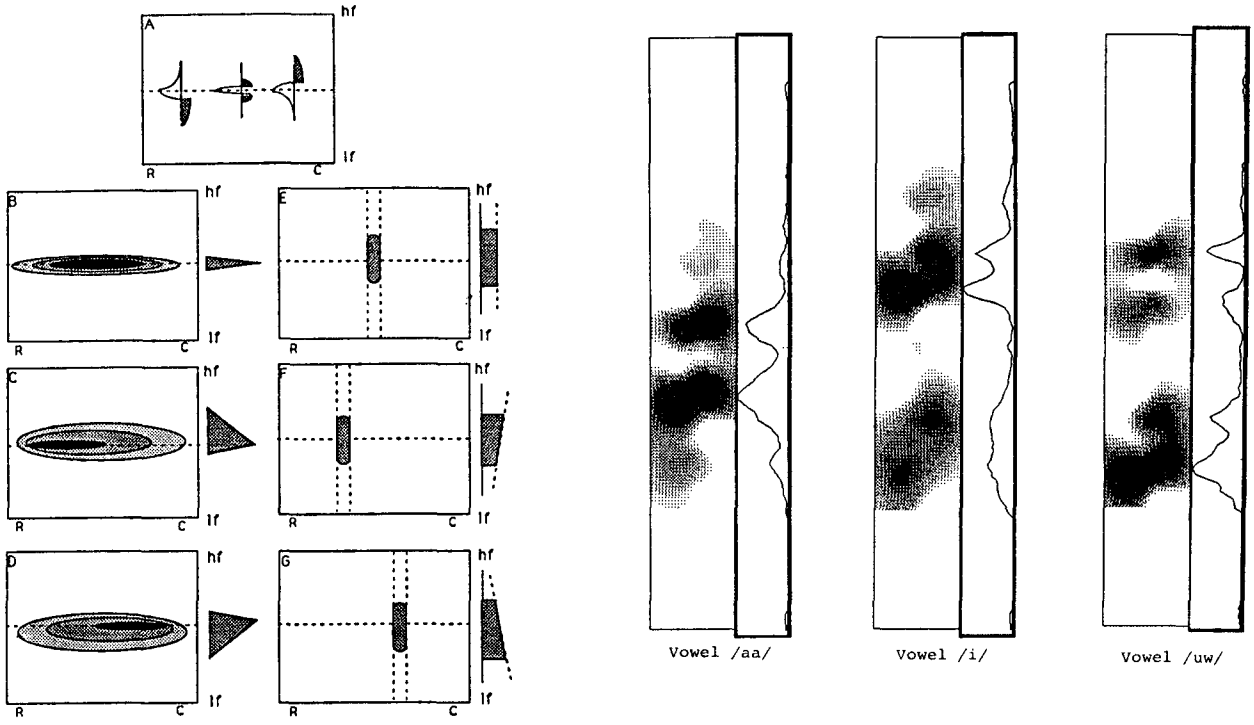

Figure 2: Spectral Representation in the Auditory Cortex: (left top) Profiles of Receptive Fields in AI Along the Iso-Frequency Planes. (left bottom) Response Patterns Elicited by Different Spectral Peaks. (right) Examples of the Distribution of Activity Produced by Speech Stimuli in a Model of AI with Spectral Orientation Columns. The Input Profiles are Shown to the Right of Each Figure.

analogous to *phase-locked loops*, where the network locks onto an incoming sequence and predicts one state ahead. An error signal between the network state and the input is fed back to control the rate of progression in the network states (Figure 3).

The system is based on a nonlinear recurrent lateral inhibitory network operating in a *hysteresis* mode which functions as a pattern generator. The network consists of a single layer of reciprocally and strongly inhibited neurons. The profile of connectivities is designed such that the patterns of the desired sequence are stable states of the network outputs. It can be shown that, when equally activated, the network settles in any one of its stable states depending on its initial conditions, i.e. displays a *hysteresis* behavior. A network generates a sequence when it cycles through its stable states. In order to control the order and rate of this process, integrating excitatory connections are formed that project from the elements of one pattern to the elements of the succeeding pattern. Only one time-constant of integration is used for all connections in the network. The varying durations of the sequence patterns are encoded not as different time constants but as different widths of the hysteresis loops between the different patterns, i.e. through the magnitudes of the inhibitory connectivities in the network.

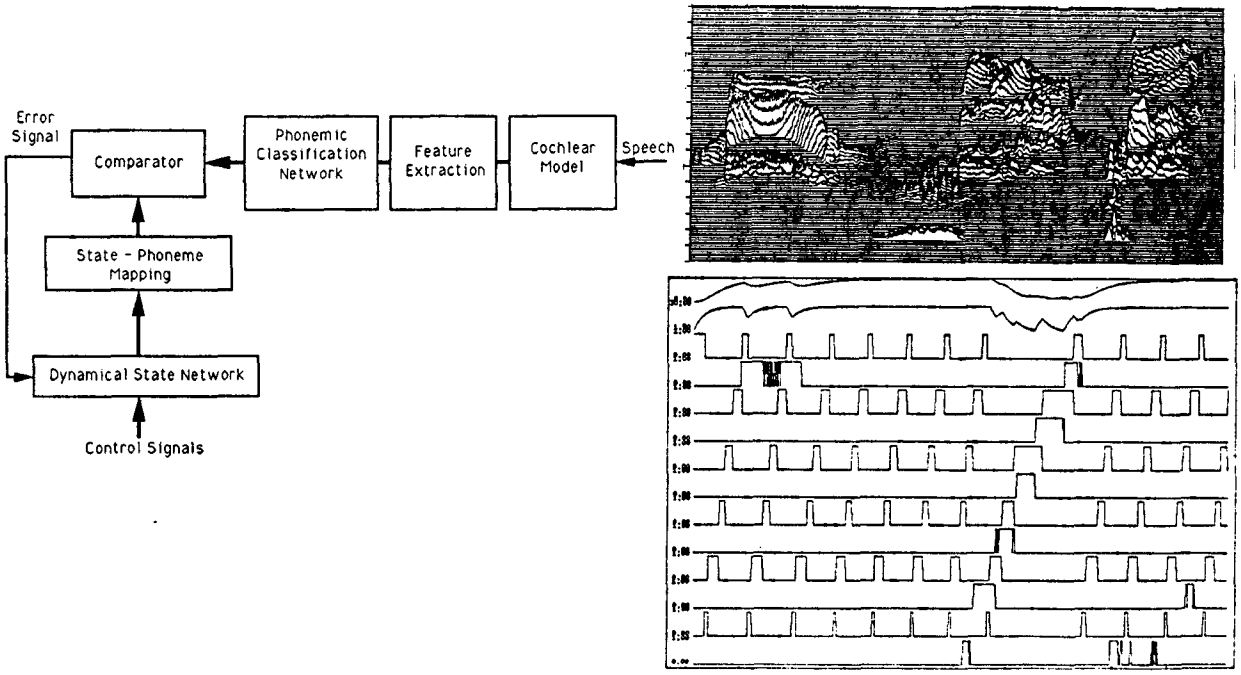

Figure 3: Temporal Sequence Recognition: (left) Network Block Diagram. (right) Correct Detection of "four" and Rejection of "nine" and "two".

The proposed network can be readily used as a recognizer of sequences applied to its input. The key concept here is the degree of correspondence between the applied input and the internally predicted state of the network. This measure is used to modulate the mode of operation in the network between a *free-cycling* mode when the correspondence is high, and an input-dominated mode when it is low. The measure is a state-dependent function derived during training, similar to a likelihood function. Thus, this measure can also be used as an indicator of the match between the applied sequence and the sequence generated by the network.

When the confidence is relatively high and the network is *free-cycling*, it automatically substitutes missing patterns and is rather insensitive to small irregularities of the input temporal durations. Therefore, in such a scheme, no time-warping is needed.

REFERENCES

[1] S. Shamma, The Acoustic Features of Speech Sounds in a Model of Auditory Processing: Vowels and Voiceless Fricatives *Journal of Phonetics* **16**, 77 (1988)

[2] S. Shamma, Spatial and Temporal Processing in Central Auditory Networks in *Methods in Neuronal Modeling*, Koch and Segev, 247 (1989)

[3] S. Shamma and K. Morrish, Synchrony Suppression in Complex Stimulus Responses of a Biophysical Model of the Cochlea *Journal of the Acoustical Society of America* **81**, 1486 (1987)

[4] S. Shamma, Spectral Orientation Columns in the Primary Auditory Cortex, University of Maryland Institute for Advanced Computer Studies Technical Report (1989)

[5] W. Byrne and S. Shamma, Dynamical Networks for Speech Recognition *Proceedings of the Annual Meeting of International Neural Network Society*, 291 (1988)