Physiologically-Motivated Synchrony-Based Processing for Robust Automatic Speech Recognition

Chanwoo Kim, Yu-Hsiang Chiu, and Richard M. Stern

Department of Electrical and Computer Engineering
and Language Technologies Institute
Carnegie Mellon University, Pittsburgh PA 15213 USA
{chanwook,ychiu,rms}@cs.cmu.edu

Abstract

This paper describes the structure and performance of a new signal processing scheme, motivated by the physiology of the peripheral auditory system, that improves speech recognition accuracy in the presence of broadband noise. An important attribute of the peripheral processing is a novel mechanism to represent the cycle-by-cycle *synchrony* in the response of low-frequency auditory-nerve fibers, in addition to the more conventional processing based on mean rate of response. It is shown that the use of the physiologically-motivated peripheral processing improves recognition accuracy in the presence of both broadband and transient noise, and that the use of the synchrony mechanism provides further improvement beyond that which is provided by the mean rate mechanism.

Index Terms: auditory modeling, robust speech recognition, auditory snchrony.

1. Introduction and Background

It has long been speculated that features based on auditory and perceptual analyses of speech encapsulate more information about speech than do the traditional generic features derived using standard signal processing techniques. The most widely adopted features based on auditory principles are the well known perceptual linear prediction (PLP) features, and even these features are outperformed by traditional MFCC features in many acoustic conditions. This, however, need not imply that our expectations (and indeed several decades of experimental observations) about the relevance of auditory processing to speech recognition were wrong. We believe instead that the mediocre performance of auditory features thus far is a consequence of both suboptimal choices of the features themselves and the lack of a good match between their characteristics and the characteristics of the speech recognition to which they are input.

This paper describes some initial results from our efforts to develop speech recognition systems based on a richer description of the peripheral auditory response to sound. In this section we review some of the previous work that has motivated our formulation. In the next section we describe our peripheral processing and feature extraction procedures in some detail. We describe in Sec. 3 some experimental results obtained using our procedures.

1.1. Early work in auditory modeling

Beginning in the early 1980s, there has been substantial interest in the use of feature sets that are developed by computational models of the auditory periphery, typically based on physiological measurements of the responses of individual fibers of the auditory nerve (*e.g.* [1, 2, 3]) Most such auditory models (sometimes called "ear models") include a set of linear bandpass filters with bandwidth that increases nonlinearly with center frequency, a nonlinear rectification stage that frequently includes short-term adaptation and lateral suppression across frequency bands, and, frequently, a more central display based on short-term temporal information. Two examples form that era include Seneff's "Generalized Synchrony Detector" (GSD) [3], and Ghitza's Ensemble Interval Histogram (EIH) mechanism [1]. The GSD describes instantaneous timing information by comparing the output of each analysis channel with itself after a delay equal to the reciprocal of the analysis frequency. The EIH mechanism estimates instantaneous frequency from the times at which the channel outputs traverse a set of fixed thresholds.

Initial evaluations of the performance of auditory models indicate that with clean speech, such approaches tend to provide recognition accuracy that is comparable to that obtained with conventional features such as MFCC or PLP parameters, and that these features can provide greater robustness with respect to environmental changes when the quality of the incoming speech decreases or differences between training and testing environments increase (*e.g.* [4]). Nevertheless, the gains in performance provided by auditory models at that time had been modest and in many cases is exceeded by the improvement in recognition accuracy provided by conventional robustness algorithms based on statistical parameter estimation (*e.g.* [5]). Furthermore, these improvements in accuracy come with great cost in computation and storage. All of these factors contributed to a decline of interest in auditory modeling for a period of time until computing resources were able to catch up with the demands of these approaches.

1.2. The role of synchrony in auditory processing

Much of our own work in this area is motivated by physiological findings by Sachs and Young [6, 7] along with similar results obtained by other research groups. Sachs and Young observed that the spectral representations of vowels developed using a measure based on the average rate of auditory-nerve response was highly dependent on signal level, while spectral representations developed from measures based on the cycle-by-cycle timing of the incoming signal maintained a high degree of consistency over a very broad dynamic range. We consider these results to be important to speech recognition because the commonly-used MFCC and PLP representations are based on precisely the same type of short-term

stimulus energy measurements that are developed by the highly non-robust physiological mean rate of response. For this reason we sought to develop a representation that reflected (at least in part) the synchronization of neural response at low frequencies to the incoming sound.

For these reasons, the potential role of auditory synchrony for feature extraction for automatic speech recognition has received increased attention in recent years. For example, a number of relatively recent studies have developed and evaluated augmentations of Seneff's GSD (*e.g.* [8, 9]). In addition, several variations of feature extraction based on *zero-crossing and peak amplitudes* (ZCPA) have appeared (*e.g.* [10, 11]), which may collectively be regarded as special cases or extensions of Ghitza's EIH model. We believe that the synchrony extraction method described below is likely to be more robust than the previously-developed algorithms because it is less dependent on the exact shape of the waveform or the specific nominal of the analysis channel. In addition, we are evaluating the systems over significantly larger databases than have generally previously been applied to systems that incorporate auditory models.

2. Speech recognition using auditory modeling

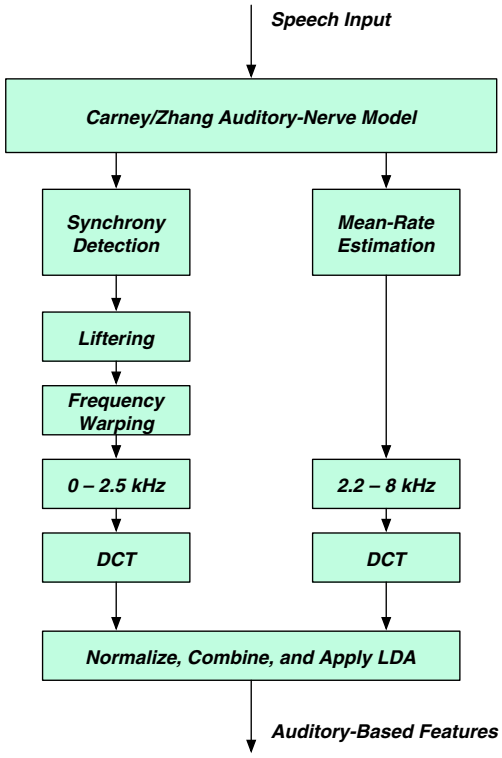

Figure 1: General description of a system that performs both mean rate and synchrony analysis. The system combines synchrony procesing at low frequencies with mean-rate processing at higher frequencies, and converts the representation into coefficients that are roughly similar to cepstal coefficients.

Our physiologically-motivated feature extraction consists of (a) a model for peripheral auditory-nerve activity, which also provides the information for mean-rate processing, (b) a mechanism for extracting synchrony information from the outputs of parallel channels of the auditory-nerve model, and (c) a mechanism to combine the mean-rate and synchrony outputs and convert them into a form that can be used by the speech recognition system. These components are summarized in broad outline in Fig. 1. We describe these components briefly in this section.

2.1. Peripheral auditory model and mean-rate analysis

The frequency analysis and subsequent nonlinear processing of sound by the peripheral auditory system establishes the representation needed for the complex separation and analyses that take place at higher levels of the auditory system. We have adopted for our purposes the implementation by Zhang *et al.* of a peripheral model developed by Carney and her colleagues, which is specified in considerable detail in the literature [12]. The model describes a number of physiological phenomena that we consider to be important such as synchrony suppression, and it is readily available in source code form on the internet.

We obtain an estimate of the mean rate of auditory-nerve output from the short-term average of the "synapse outputs" of the Carney model, which are continuous functions that are proportional to the instantaneous rate of auditory-nerve firings in a given frequency band. Because the peripheral auditory model is highly nonlinear, the amplitude of incoming utterances is adjusted on a sentence-by-sentence to maintain a constant value of the power of the total signal in the frequency range 0 to 4 kHz, excluding silence regions.

2.2. Synchrony extraction

We have explored several ways of extracting synchrony information for speech recognition. The processing in this paper attempts to extract synchrony in a way that reflects the frequency content of the original signal, rather than merely the center frequency of each analysis channel. We first pass the output of each channel of the auditory model through a second bandpass filter with the same frequency response as the auditory filter for that channel to reduce the harmonic distortion introduced by nonlinearities in the peripheral auditory processing. The short-time Fourier transform of the outputs of the bandpass filters is computed, and these frequency responses are averaged across channels. This produces a high-resolution spectral representation at low frequencies for which the auditory nerve is synchronized to the input up to about 2.2 kHz, and which includes the effects of all of the nonlinearities of the peripheral processing.

We remove the horizontal striations typically seen in narrowband spectrograms (which reflect the pitch of the incoming signal) by applying a discrete-cosine transform (DCT) to the frequency response, applying a short-time "lifter" to the inverse transform, and then returning to the frequency domain using an inverse DCT.

2.3. Development of features for speech recognition

The features used for speech recognition are developed by merging the synchrony outputs at low frequencies with the mean rate outputs at higher frequencies. First the synchrony outputs, which emerge initially as a linear function of frequency, are warped along the frequency axis so that they while the mean rate outputs are warped along the frequency axis in order to correspond to the nonlinear dependence on frequency of the center frequencies of the

analysis channels used to develop the mean rate outputs.

Components from the frequency-warped synchrony outputs between approximately 0 and 2.5 kHz and components from the mean rate outputs between approximately 2.2 to 8 kHz are preserved for further processing. These outputs are subjected to a final DCT which produces a set of coefficients that are somewhat similar to cepstral coefficients. The present implementation uses the first 8 DCT coefficients derived from the synchrony outputs and the first 5 DCT coefficients derived from the mean rate outputs. These are concatenated into a vector of 13 coefficients which serve as the basic input to the speech recognition system. Delta and delta-delta coefficients are obtained in the same fashion as is normally done for the CMU SPHINX system.

Fig. 2 compares the outputs of conventional MFCC processing and the processing developed in this paper. The upper panel is a spectrogram of an utterance from the DARPA Resource Management (RM) corpus that had been corrupted by white noise at a signal-to-noise ratio (SNR) of +10 dB. The center panel is a reconstructed spectrographic representation of the features that are developed by conventional MFCC processing with cepstral mean normalization. The lower panel is a similar representation of the combined mean rate and synchrony auditory-model outputs. We note that the output of the auditory model provides a clearer picture that suppresses many of the effects of the noise.

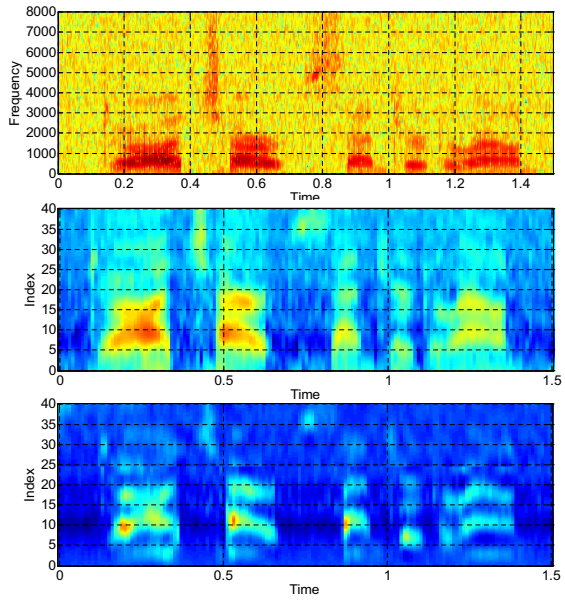

Figure 2: Comparison of outputs of conventional and physiologically-motivated signal processing. Panels from top to bottom depict (a) a wideband spectrogram of an RM Utterance corrupted by white noise at an SNR of +10 dB, and reconstructed spectrograms developed from (b) conventional MFCC processing with cepstral mean normalization, and (c) the combined mean rate and synchrony auditory-model outputs.

3. Experimental results

The feature extraction schemes described above were evaluated by comparing the recognition accuracy obtained with the CMU SPHINX-III system using conventional MFCC processing, the mean rate outputs alone of the auditory model, and the combined mean rate and synchrony outputs. Two standard speech corpora were used for these evaluations. The first corpus consisted of a subset of 1600 training utterances and 600 testing utterances from the DARPA Resource Management (RM) database. The second larger corpus was the DARPA Wall Street Journal WSJ0 (WSJ) database, which consisted of 7024 training sentences and 651 testing sentences. Since we are concerned primarily with the relative performance of the various signal processing schemes considered, no attempt was made to fine tune the parameters of the SPHINX trainer and decoder to minimize the absolute error rate.

Experimental results are shown in Fig. 3 for the RM task and in Fig. 4 for the WSJ task. In each case the speech is corrupted by both white noise (upper panels) and by musical segments of the DARPA Hub 4 Broadcast News database (lower panels). Word accuracy (defined as 100% minus the word error rate as defined by NIST) is plotted as a function of SNR for systems using conventional MFCC features (squares), auditory processing using mean rate information only (triangles), and the auditory processing using the combination of mean rate and synchrony information (diamonds) as depicted in Fig. 1.

It can be seen that the use of the physiologically-motivated signal processing results in a substantial improvement in word accuracy, particular at SNRs in the range of 5 to 15 decibels. We prefer to characterize improvement as the amount of threshold shift provided by the processing (as opposed to the percent improvement at a particular SNR). The combined processing that includes synchrony provides an improvement of about 15 dB SNR in white noise for the RM task and perhaps 10 dB for the WSJ task. Improvements in the presence of background music are far more limited, about 3-4 dB. Processing using mean rate alone is usually not as effective as processing that includes synchrony. While we believe that the worse performance observed with background music is a partly a result of suboptimal input signal normalization, we have also observed a similar effect using more traditional signal processing for robust speech recognition (*cf.* [13]).

Figure 5 compares the recognition accuracy obtained using mean-rate processing of the outputs of the peripheral auditory model of Zhang *et al.* [12] and similar procesisng using a simplified peripheral model. The simplified model consists of a cascade of (1) a bank of fourth-order gammatone filters with the same center frequencies as those of each channel of the Zhang emphet al. model, (2) a full-wave rectifier of each of the channel outputs, and (3) a memoryless compressive nonlinearity that has the same input-output characteristics as the corresponding long-term response of the Zhang *et al.* model at each center frequency. It is clear that the detailed processing of the Zhang *et al.* model provides a representation that is more robust in noise than the simplified model, even though we do yet understand on a deep level exactly which aspects of the Zhang *et al.* model are the most instrumental in accomplishing this.

4. Summary

We have developed a new approach to the representation of synchrony information at the level of the auditory nerve which has led to substantially improved speech recognition accuracy compared to both conventional cepstral processing and compared to similar peripheral that is based on the mean rate of response only.

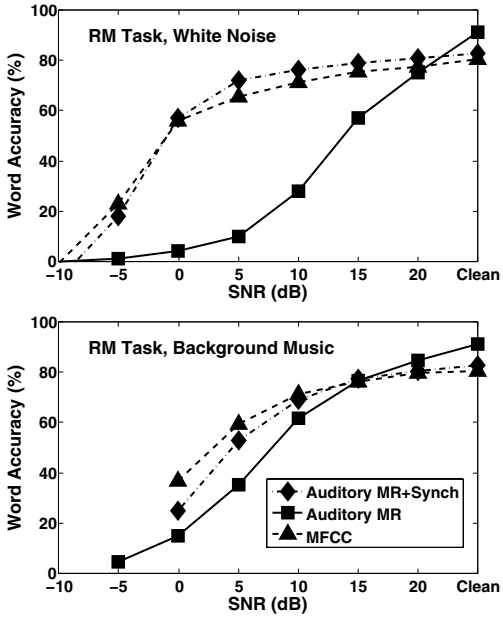

Figure 3: Recognition accuracy (100% minus word error rate) for the DARPA Resource Management (RM) task. Plotted are results using baseline MFCC coefficients with cepstral mean normalization (triangles), the auditory model using mean rate only (squares), and the auditory model with mean rate and synchrony (diamonds).

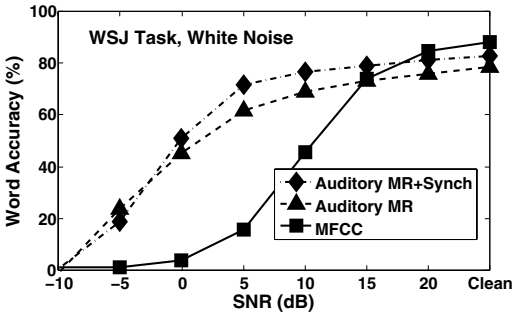

Figure 4: Same as Fig. 3, but results are ploted for the DARPA Wall Street Journal Dictation (WSJ) task.

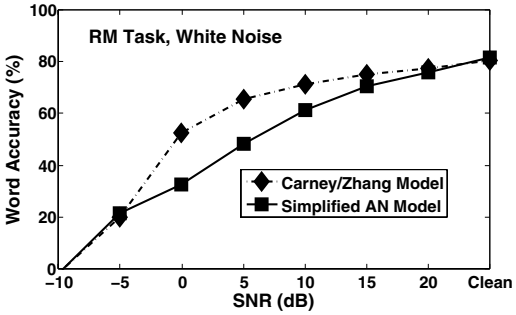

Figure 5: Comparison of mean-rate results using the model of Zhang *et al.* (diamonds), and a simplified auditory-nerve model (squares).

5. Acknowledgements

This research was supported by the National Science Foundation (Grant IIS-0420866) and the DARPA GALE Project. We are also indebted to Evandro Gouvea and Balakrishnan Narayanaswamy for their contributions to this research.

6. References

[1] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 2, pp. 52–59, 1986.

[2] R. F. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, May 1982, pp. 1282–1285.

[3] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 15, pp. 55–76, 1988.

[4] H. Meng and V. W. Zue, "A comparative study of acoustic representations of speech for vowel classification using multi-layer perceptrons," in *Proceedings of the International Conference of Spoken Language Processing*, 1990.

[5] Y. Ohshima and R. M. Stern, "Environmental robustness in automatic speech recognition using physiologically-motivated signal processing," in *Proceedings of the International Conference of Spoken Language Processing*, 1994.

[6] M. B. Sachs and E. D. Young, "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate," *J. Acoust. Soc. Amer.*, vol. 66, pp. 470–479, 1979.

[7] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1381–1403, 1979.

[8] P. Cosi, "Auditory modeling for speech analysis and recognition," in *Visual Representation of Speech Signals*, M. Cooke, S. Beet, and M. Crawford, Eds., 1992, pp. 205–212.

[9] A. M. A. Ali, J. Van Der Spiegel, and P. Mueller, "Robust auditory-based speech processing using the average localized synchrony detection," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 279–292, 2002.

[10] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 55–69, 1999.

[11] M. Ghulam, J. Horikawa, and T. Nitta, "A pitch-synchronous peak-amplitude based feature extraction method for robust asr," *Proc. Int. Conf. Spoken Lang. Processing*, vol. I, pp. 517–520, 2005.

[12] X. Zhang, M. G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenologial model for the response of auditory-nerve fibers: I. nonlinear tuning with compression and suppresion," *J. Acoust. Soc. Amer.*, vol. 109, pp. 648–670, 2001.

[13] B. Raj, V. N. Parikh, and R. M. Stern, "The effects of background music on speech recognition accuracy," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1997, vol. 2, pp. 851–854.