interspeech 2008 incorporating SST 08

Analysis of Physiologically-Motivated Signal Processing for Robust Speech Recognition

Yu-Hsiang Bosco Chiu and Richard M Stern

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

Abstract

This paper discusses the relative impact that different stages of a popular auditory model have on improving the accuracy of automatic speech recognition in the presence of additive noise. Recognition accuracy is measured using the CMU SPHINX-III speech recognition system, and the DARPA Resource Management speech corpus for training and testing. It is shown that feature extraction based on auditory processing provides better performance in the presence of additive background noise than traditional MFCC processing and it is argued that an expansive nonlinearity in the auditory model contributes the most to noise robustness.

Index Terms: auditory modeling, robust speech recognition, auditory analysis

1. Introduction

The use of automatic speech recognition (ASR) systems is presently extending to a wide variety of application areas including task-oriented dialog systems, meeting transcription, and telematic assistance. Robustness to environmental and acoustical change is an increasingly important issue. Motivated by the human ability to recognize speech under adverse environments, feature extraction based on auditory physiology has been applied to ASR with some success over a period of decades (*e.g.* [1-9]). In addition, the log scale frequency analysis and amplitude compression that are major components of auditory models are important components of conventional feature extraction schemes such as mel-frequency cepstral coefficients (MFCC) [10], and perceptual linear prediction (PLP) [11].

Although the conventional MFCC and PLP methods for feature extraction function quite well when acoustical environments for training and testing are matched, their performance degrades seriously when they are applied in noisy environments especially when training and testing conditions are mismatched. A number of feature extraction methods that are motivated by results from auditory physiology have been developed over the years, which have yielded systems that outperform traditional approaches such as MFCC or PLP in the presence of noise and other adverse conditions [5-9].

In this paper, we first describe the feature extraction scheme used, which is based on an implementation of the detailed model of the auditory periphery by Seneff [2]. We then discuss the impact of each stage of the auditory model on speech recognition accuracy. Similar analyses have been performed in [5, 6]. Ohshima and Stern [5] considered only the short-term adaptation, lowpass filter, and automatic gain control (AGC) stages, which are not critical in our analysis. Tchorz and Kollmeier [6] concluded that the adaptive compression stage of the auditory model (which corresponds to the hair cell model of the Seneff model) is of the greatest importance. We elaborate on this issue in the present paper.

In Sec. 2 we review some of the previous work that has motivated our formulation and system implementation. The extracted features are evaluated and a more detailed analysis of the robustness contribution of each stage of the auditory model is discussed in Sec 3. Finally in Sec. 4, we support our assertion that the auditory nonlinearity is of paramount importance by applying it to a conventional log Mel spectrum.

2. Background

In general, to extract features from an incoming speech signal for speech recognition, the incoming speech is segmented into short time segments and these segments are analyzed to reveal their frequency characteristics while preserving the time-varying characteristics inherent in the signal.

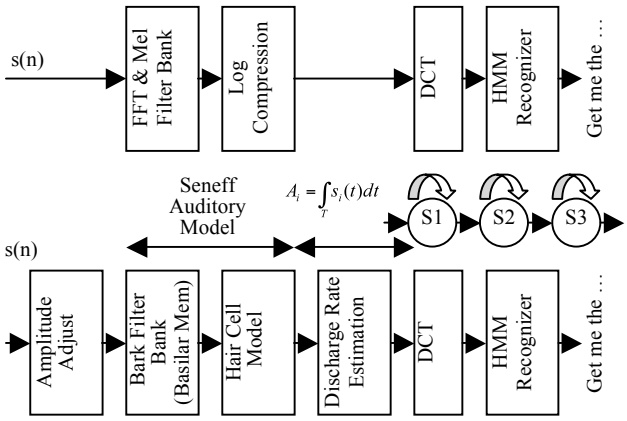

Figure 1 *Block diagram of traditional MFCC processing* (*upper panel*) *compared with the Seneff auditory-based speech processing system* (*lower panel*).

2.1. Feature extraction in the auditory periphery

Generally speaking, models of auditory-based feature extraction can be divided into two main stages. The first stage is the model of the auditory periphery, for which we adopt the implementation of Seneff [2] to deal with sound transformations occurring in the early stages of the hearing process. The second stage is a series of operations intended to convert the auditory outputs into estimates of short-term average firing rate, and subsequently into features that are like cepstral coefficients. Fig. 1 summarizes this processing (lower panel) and compares it to conventional MFCC processing (upper panel). The Seneff auditory model is expanded in the block diagram in Fig. 2.

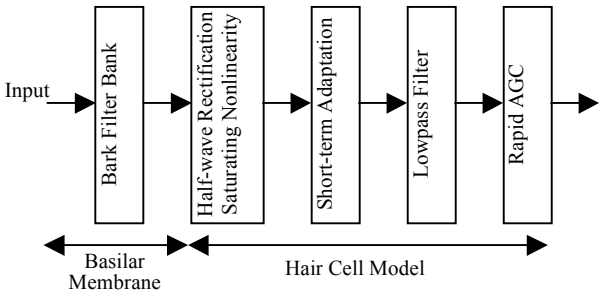

Figure 2 *Detailed structure the Seneff auditory model.*

2.1.1. Basilar Membrane

After amplitude adjustment such that the maximum amplitude of the input signal normalized to 1, the speech signal is passed through a Bark-scaled filter bank of 40 bandpass filters representing the frequency analysis by the basilar membrane in the cochlea. The bandwidth of the filters is designed to mimic human frequency resolution with relatively narrow-band filters in the low-frequency region and wider-band filters in the high frequency region.

2.1.2. Hair Cell Model

Seneff's hair cell model attempts to capture the electrochemical transformation from basilar membrane vibration, represented by the output of the filter bank, to the time-varying neural firing rate of each fiber. It consists of several stages: (a) halfwave rectification with a compressive nonlinearity, to represent the inherently positive nature of the rate of spike generation and the input-output relationships between amplitude and spike rate, which is referred to here as the rate-level function (b) short-term adaptation, which models certain aspects of the electrochemical spike generation process, (c) a lowpass filter, which represents the loss of detailed timing information at higher frequencies, and (d) a rapid automatic gain control (AGC) which represents, among other attributes, the limit on spike rate imposed by the inability to generate spikes in short succession. The panels of Fig. 3 illustrate the response of the system to a tone burst at 2000 Hz after the initial bandpass filtering, after the initial rectification and saturation, after the initial adaptation, and after the AGC, respectively.

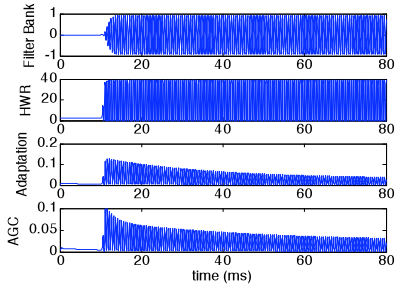

Figure 3 *Output of each intermediate stage in the Seneff inner hair cell model in response to a 2-kHz input signal.*

2.1.3. Discharge Rate Estimation

As observed from neural recordings in physiological experiments, we could describe the sound representation in higher stage of auditory system by the number of firings within a short time interval in its response to sound stimuli, as it is proportional to the loudness of the sound stimuli (*e.g.* [12]). When the input stimulus is kept at an appropriate level to avoid saturation in the auditory nerve fibers, the "firing pattern" characterized by the number of firings could well preserve the frequency content and describe how sound is represented in higher stages of the auditory system in the human brain. Since the outputs of the auditory model are measured in spikes/second, we consider the discharge rate to be described by the number of spikes within a certain time interval. We integrated these outputs over a 20-ms frame because that duration is widely used for automatic speech recognition:

$$A_i = \int_T s_i(t)dt \qquad \text{for i} = 1,2,...,\text{N} \qquad (1)$$

where N is the number of channels. For a speech frame at time n, the corresponding feature coefficients are computed by the DCT of the channel outputs as in MFCC processing to reduce the dimension and obtain the final features.

3. Experimental Results

3.1. Performance compared with MFCC processing

The feature extraction scheme described above was applied to the DARPA Resource Management (RM) database. This database contains Naval queries with 1600 training utterances and 600 testing utterances (72 speakers in the training set and another 40 speakers in the testing set representing a variety of American dialects). To evaluate the performance under noise, white noise from NOISEX-92 was artificially added to the testing set with energy adjusted according to a pre-specified noise level (with SNRs of 0 dB, 5 dB, 10 dB, 15 dB, 20 dB). We used CMU's SPHINX-III speech recognition system. Cepstral-like coefficients were obtained for the auditory model by computing the DCT of the outputs of the estimator of discharge rate in each frequency band, as in the lower panel of Fig. 1. Seven such coefficients were obtained for each frame in the auditory model, compared to thirteen cepstral coefficients for traditional MFCC processing. Cepstral mean normalization (CMN) was applied in both cases. A comparison of speech recognition accuracy obtained with the auditory model (defined as 100% minus the word error rate (WER)) with the accuracy obtained using traditional MFCC processing is shown in Fig. 4.

As can be seen from Fig. 4, speech recognition accuracy in the presence of background noise is greater when the auditory model is used than the accuracy obtained using traditional MFCC processing, especially around the 10-dB noise level (around a 7 dB improvement over MFCC). While we have previously obtained much better results using the Zhang/Carney auditory model [8] [13], we use the Seneff model at present because its structural simplicity facilitates stage-by-stage analysis. Next, we consider feature extraction at different stages of the auditory model output to determine which component has the greatest impact on recognition accuracy.

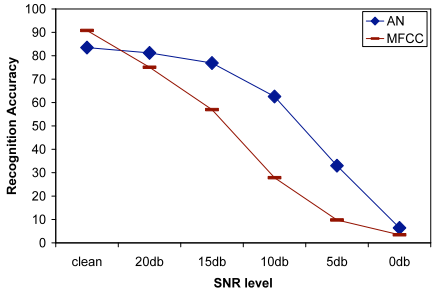

Figure 4 *Comparison of the percentage recognition accuracy (100% minus the word error rate) using features based on auditory processing (diamonds) and MFCC processing (short lines) for the DARPA Resource Management (RM) database.*

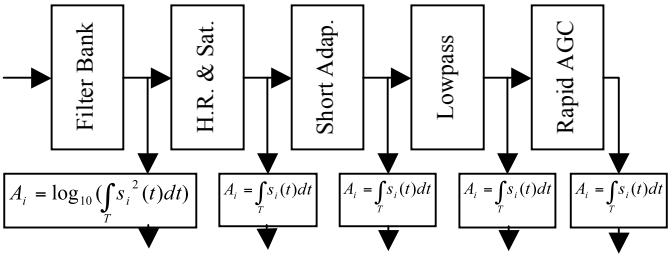

Figure 5 *Features extracted from each stage of the auditory model.*

3.2. Significance of each stage of the auditory model

To understand why using auditory processing could give us such improvement in the presence of noise, it is helpful to evaluate the contribution of each of its stages. Since the auditory model is fine tuned to the physiological data and each stage depends on appropriate input from the previous stage, taking out any stage is likely to cause the system to malfunction and its effect will be unable to be analyzed appropriately. To analyze the effect of each stage while maintaining the functionality of the auditory model, we compared the performance of each of stage after the filter bank by integrating its output over 20 ms as in Fig. 5. The sole exception is the filter bank output which was obtained by calculating the short-term energy of each bandpass filter output, taking the log, and computing the DCT, in a fashion similar to that of traditional MFCC processing. These results evaluated on the RM database using the SPHINX-III speech recognition system are shown in Fig. 6 and discussed in the following paragraphs.

3.2.1. Effect of the rectification and nonlinearities

To evaluate the effect of the rate level function, we first compare the recognition performance with features extracted before and after the half-wave rectification/saturating nonlinearity stage. As can be seen from Fig. 6, extracting features directly from the outputs of the filter bank (circles) provides performance that is quite similar to the result of MFCC processing (short lines). This result is somewhat expected as both are based on similar concepts (the filter bank simulates the frequency resolution of human ear while the log operation simulates the loudness curve). On the other hand, if we compare the result of features extracted from the outputs of the rectification/saturating nonlinearity stage (crosses) with the result of the filter bank outputs, the performance is much improved under noisy condition while somewhat degraded under clean speech.

As shown in Fig. 7, the rate level function functions as a soft clipping mechanism, which limits both small and large amplitudes of sound. Because small-amplitude sounds are more easily affected by noise, this mechanism could help reduce the noise degradation. For example, as shown in the lower panel, which depicts the amplitude histogram of clean speech in the training data, under certain noise levels, such as –60 dB, speech signals with large amplitude such as –40 dB will only be slightly affected by additive noise after compression. In contrast, speech signals with small amplitudes such as –80 dB (close to the silence region), additive noise of –60 dB is 10 times larger than clean speech and causes huge amount of degradation after compression. Attenuating the waveform during small-amplitude segments of sound can help reduce the degradation caused by noise, but the resultant deliberate signal distortion can degrade recognition accuracy for clean speech.

3.2.2. Effect of short term adaptation

As in the previous stages, we can assess the effect of short term adaptation by comparing results obtained from features derived from the outputs of the half wave rectifiers (crosses) and the outputs of short term adaptation (triangles). These are the inputs and outputs of the short-term adaptation stage of the auditory model. The transient enhancement produced by the short-term adaptation improves recognition accuracy only slightly, as seen in Fig. 6. This finding is somewhat different from the conclusions in [6] and [9]. Our implementation includes both integration (which is lowpass in nature with a cutoff frequency around 50 Hz) and CMN (which is highpass, removing the DC component). The net effect of these modules is that of a bandpass filter which emphasizes the low frequencies that are most significant in modulation-spectrum analyses. This may limit the potential benefit of short-term adaptation, which is believed by at least some researchers (*e.g.* [6], [9]) to have a similar effect on the incoming signal.

3.2.3. Effect of the lowpass filter

For the third step, we compare features directly from the short-term adaptation stage output with features from the outputs of the lowpass filters to examine the effect of the lowpass filter stage in auditory model. As shown in Fig. 6 (triangles versus squares), the presence of the lowpass filter has little effect on the results obtained. This is somewhat as one would expect, as the feature extraction includes integration over the output, which could also be seen as a kind of lowpass filtering. Since the cutoff frequency of the lowpass filter stage (around 4 kHz) is much greater than the cutoff frequency of integration (around 50 Hz for a 20-ms period), the removal of the lowpass filter here will not have much effect on performance.

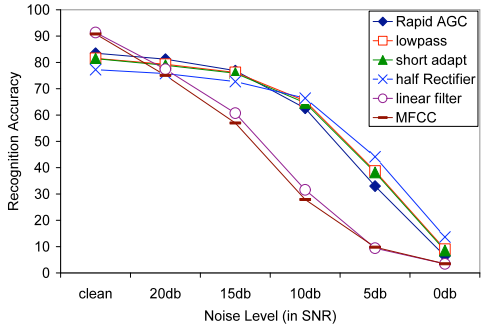

Figure 6 *Comparison of recognition accuracy for the RM database using features extracted from outputs of each stage of auditory model. (See legend for details.)*

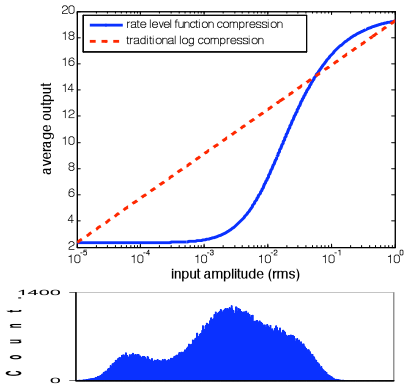

Figure 7 *Upper panel: Rate level function (line) in the half wave rectification stage compared with traditional log compression (dots). Lower panel: magnitude (rms) histogram for clean speech.*

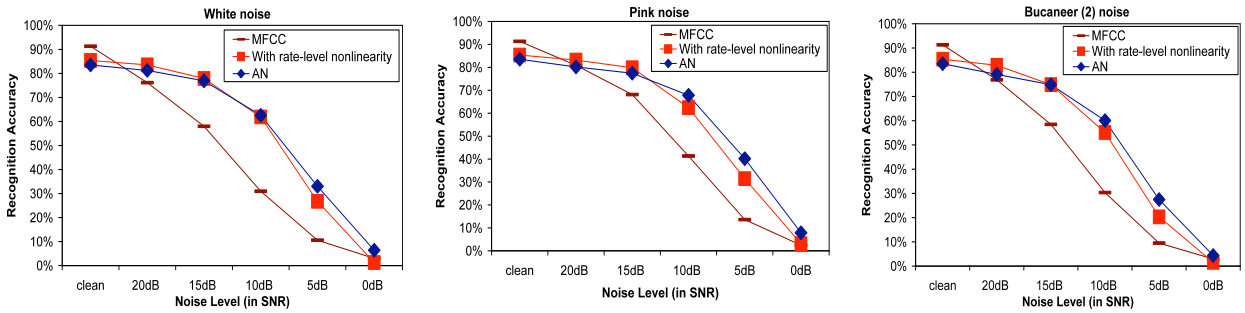

Figure 8 *Comparison of recognition accuracy for the RM database obtained by applying the auditory rate-level nonlinearity directly to log Mel spectral values (squares), with traditional MFCC processing (short line) and with whole auditory processing (diamonds).*

3.2.4. Effect of AGC

Because the effect of the AGC is similar to that of short-term adaptation (as can be seen in Fig. 3), recognition accuracy is slightly improved for clean speech due to transient enhancement, compared to the results obtained directly from the lowpass filter output before the final AGC stage (squares and diamonds in Fig. 6).

4. Application of auditory nonlinearity to log Mel spectral coefficients

We argued in Sec. 3.2.1 that the most important aspect of the auditory model was the nonlinearity associated with the hair cell model. To the extent that this is true, we should be able to obtain a similar benefit by applying such nonlinearity to conventional MFCC-like feature extraction. Toward this end we interposed the logit function in the upper panel of Fig. 7 between the log of the triangularly-weighted frequency response and the subsequent DCT operation in traditional MFCC processing. Results in Fig. 8 for speech in the presence of white noise, pink noise, and "buccaneer" noise from the NOISEX-92 database show a similar improvement in recognition accuracy seen in Fig. 6, corresponding to about 7-dB improvement around the 10-dB white noise level. In other words, the benefit of the auditory nonlinearity can be obtained without incurring the computational complexity associated with other aspects of auditory modeling, at least to some extent.

5. Conclusions

We have examined the relative effectiveness of the various stages of the model of the auditory periphery proposed by Seneff for improving the recognition accuracy of speech in the presence of broadband noise. Detailed robustness contributions from each stage of auditory model are also described and discussed. Results obtained using the DARPA Resource Management database with CMU's SPHINX-III recognition system indicate an improvement of about 7 dB for the Seneff model for these maskers. We also found that the saturating nonlinearity contributes the most to robustness at lower SNRs while transient enhancement in the rapid AGC and short term adaptation, on the other hand, enhance recognition accuracy only for clean speech. By applying the same nonlinearity to the log Mel spectrum, one can achieve similar results with conventional MFCC processing.

6. Acknowledgements

This work was supported in part by the National Science Foundation (Grant IIS-0420866).

7. References

[1] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," *Proc. IEEE Int. Conf. on Acoust. Speech, and Sig. Proc. (ICASSP),* May, 1982.

[2] S. Seneff, "A joint synchrony/mean rate model of Auditory Speech Processing," *J. Phonetics*, **16**: 55-76, 1988.

[3] O. Ghitza, "Temporal Non-place Information in the Auditory-nerve Firing Patterns as a Front-End for Speech Recognition in a Noisy Environment," *J. Phonetics*, **16**:109-123, 1988.

[4] J. R. Cohen, "Application of an auditory model to speech recognition," *J. Acoust. Soc. Amer.*, **85:**. 2623-2629, 1989.

[5] Y. Ohshima and R. Stern, "Environmental Robustness in Automatic Speech Recognition Using Physiologically-Motivated Signal Processing," *Proc. Int. Conf. on Spoken Language Proc. (ICSLP)*, September, 1994.

[6] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Amer*., **106:**2040–2050, 1999.

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

[8] C. Kim, Y.-H. Chiu, and R. Stern "Physiologically-Motivated Synchrony-Based Processing for Robust Automatic Speech Recognition," *Proc. ICSLP*, September 2006.

[9] M. Holmberg, D. Gelbart and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing" *IEEE Trans. on Speech and Audio Processing*, vol. 14, pp. 43-49, 2006

[10] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE *Trans. Acoust., Speech, Signal Processing*, **28:**357-366,1980.

[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, **87:**1738-1752, 1990.

[12] 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.*, **66:**470-479, 1979.

[13] X. Zhang, M.G. Heinz, I. C. Bruce, and L. H. Carney, "A phenomenological model for the response of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression," *J. Acoust. Soc. Amer.* **109:**648-670, 2001.