Internal report, Dept. of EE & CE, The University of Newcastle, Australia, 12 May 1996

A REVIEW OF THE PERFORMANCE OF AUDITORY PROCESSING FRONT ENDS FOR AN AUTOMATIC SPEECH RECOGNIZER IN ADVERSE ENVIRONMENT

Beng T. TAN, Minyue Fu

Dept. of Electrical and Computer Engineering,
The University of Newcastle,
NSW 2308, Australia.
beng@ascod.newcastle.edu.au

ABSTRACT

We review briefly the the performance of auditory processing front ends for an automatic speech recognizer (ASR) in adverse environment. At low SNR, the improvement of auditory model front ends can up to 20 % for speech degraded by white noise and 4 % for real world noise. They are particularly insensitive to broadband spectral distortion.

1. INTRODUCTION

The auditory model front ends have been applied widely in speech technology. However, it is still arguable whether an auditory model front end improves the performance of ASR with undegraded speech. Some experimental results show that auditory model front ends are either compatible to or not perform as well as a mel scale filter bank (MFB) or short time Fourier transform (STFT) front ends with undegraded speech [2, 6]. On the other hand, some results support that the auditory model front ends help to improve the performance of ASR and the improvement ranges from 1 % to 5 % [4, 5].

Under the adverse environment, the results are more consistent and more encouraging. It has been shown by a range of experiments [9, 1, 2, 5, 6, 8] that auditory model front ends are more robust than the conventional analysis techniques under adverse conditions. However, considering the heavy computational load of auditory model front ends, we would like to know how much improvement can we gain by using an auditory model front end in adverse conditions?

It is always difficult to compare the performance of speech recognizer across different systems. Most experiments are comparing the auditory model front end against either a mel scale filter bank (MFB) or a short time Fourier transform (STFT) front ends. For these two front ends, LPC or cepstral analysis can be used to reduce the dimension of the final acoustic feature vector. In this review, we broadly categorize the front ends into auditory models, MFB and STFT. Three adverse conditions are studied, i.e. white noise, real world noise and spectral tilt. The percentage given here is just a rough estimate or average of the experimental results. All systems are trained with clean speech and tested on degraded speech.

2. NOISE

Noise was added to utterances at various signal-to-noise-ratio (SNR). Each SNR was measured by calculating the ratio of speech energy to noise energy. Energy levels were measured by averaging the square of the signal across the entire utterance.

2.1. White Noise

The simplest form of an artificial noise is white noise. It has been used in most system to test the performance of a speech recognizer in a noisy condition.

In a speaker dependent isolated word DTW based recognition system [2], Ensemble Interval Histogram (EIH) auditory model are compared against STFT front end. With increasing levels of background noise, the recognition scores based on the Fourier spectrum decline more markedly than those using the EIH front end. The decline in the STFT performance at lower SNRs is particularly marked when the signals were spoken by males. For male speaker, the different in recognition score between the EIH and STFT front ends range from 10-15 % (at 24 dB SNR) to 30-40 % (at 12 dB SNR). For female speaker, the different only become significant at 18 dB SNR or lower. The EIH auditory model is about 10 % better than the STFT front end when the signals were spoken by females. In most cases, the error rates increase by about 40-60 % at 12 dB SNR than at ∞ dB.

Another experiment comparing a modified Seneff model and MFB front ends are performed at 15 dB SNR. The increase of recognition score is about 25 % [5].

2.2. Real World Noise

The database of recorded noise is collected by placing a microphone in a common area where many people congregate and carry on conversations. Recorded noise can simulate real-world noise conditions more accurately.

In an isolated word recognition experiment [6], a recorded noise was added to clean speech. The auditory used in

this study includes Seneff auditory model [7] and EIH auditory model [3]. For clean speech and speech with 30 dB SNR, auditory models perform similarly to MFB cepstral front end. Below 30 dB SNR. the auditory models perform slightly better than MFB cepstral front end. The difference between the auditory models and the MFB cepstral front ends range from 0.6 % (at 24 dB SNR) to 4 % (at 6 dB SNR). Although auditory models are 4 % better than the MFB cepstral front end at 6 dB SNR, the error rate increases by about 22 %, making the usefulness of the system at this conditions questionable.

3. SPECTRAL TILT

Auditory model front end are significantly better than MFB front ends under spectral tilt conditions. When 6 dB/octave spectral tilt is applied to speech signal [5], the auditory model front ends are about 60 % better than the MBF front end. The degration of recognition score for auditory model front end in this condition is small.

The stability of the auditory models front ends against broadband spectral distortion is also observed in [8]. The auditory spectrum are compared with linear power spectrum with different preemphasis factors. The auditory spectrum remains relatively stable, with the preemphasis effects mostly concentrated in the high frequency channels.

4. CONCLUSION

In noise free condition, the performance of auditory models front ends are compatible to the MFB and STFT front end. In the adverse conditions, the auditory model front ends has the ability to suppress the noise and they are not sensitive to spectral distortion. These two characteristics are contributed mainly by the nonlinear stage (after the cochlea filters) of the auditory models.

When white noise is added to speech signal, the improvement of auditory model front ends over both the MFB and STFT front ends is more than 20 % at SNR lower than 30 dB. For real world noise, the improvement is in the order of a few percentage points only.

Although auditory model front end out perform the other two front ends at low SNR, the recognition score degrades significantly (up to 20 %) at low SNR. Therefore, it is questionable whether under this condition the system is still useful.

The insensitivity of the auditory model against broadband spectral distortion indicate that the preemphasis filter may not be important here. In addition, the auditory model front ends may provide robustness in real application when the characteristics of the input channel vary.

5. REFERENCES

[1] M. D. Chau and C. D. Summerfield. Auditory models as front-ends for speech recognition in high noise environments. In J. Pittam, editor, *Proc. of the Fourth Australian Int. Conf. on Speech Science and Technology*, pages 625–628. Australian Speech Science and Technology Association, 1992.

[2] 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.

[3] O. Ghitza. Auditory nerve representation as a basis for speech processing. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 453–486. New York: Marcel Dekker, 1992.

[4] M. J. Hunt and C. Lefèbvre. Speech recognition using a cochlear model. pages 1979–1982, 1986.

[5] M. J. Hunt and C. Lefèbvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 262–265, 1989.

[6] C. R. Jankowski, H.-D. H. Vo, and R. P. Lippmann. A comparison of signal processing front ends for automatic word recognition. *IEEE Trans. Speech and Audio Proccessing*, 3(4):286–293, 1995.

[7] S. Seneff. A computational model for the peripheral auditory system: Application to speech recognition research. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 37.8.1–37.8.4, 1986.

[8] K. Wang and S. Shamma. Self-normalization and noise-robustness in early auditory representations. *IEEE Trans. Speech and Audio Proccessing*, 2(3):421–435, 1994.

[9] D. Woo, P. Dermody, R. Lyon, and B. Lowerre. Auditory model interfaces into a DTW recognizer. In J. Pittam, editor, *Proc. of the 4th Australian International Conference on Speech Science and Technology*, pages 784–788, 1992.