DUAL-CHANNEL AUDITORY SPECTRUM MODELING

*Jayadev Billa**

Department of Electrical Engineering
University of Pittsburgh
Pittsburgh, PA 15261.
jbilla+@pitt.edu

ABSTRACT

In this paper we propose a new approach to the modeling of speech based on cues from the peripheral auditory system. Our approach attempts to incorporate the dynamic adaptation of biological auditory systems to varying sound by simplistically formulating a dual-processing strategy that treats unvoiced and voiced speech as deserving of different processing. Preliminary studies show that this approach possesses significant noise robustness.

1. INTRODUCTION

Humans are able to accurately recognize speech over a wide variety of acoustic environments while automatic speech recognition (ASR) systems show degradation in performance even under mild variations in the acoustic environments used in training and testing the systems (e.g., Acero, 1993[1]).

While traditional signal processing approaches have been vigorously pursued, the success of techniques that use crude perceptual cues, such as in mel-frequency warping (e.g., Davis and Mermelstein, 1980[2]), perceptual linear prediction (PLP)[3], RASTA processing of speech[4] and sub-band correlation (SBCOR) analysis [5, 6], provide clear evidence that cues from studies of the peripheral auditory system can be successfully and meaningfully integrated into ASR front-ends, thus providing better and more robust representations of speech. Attempts at using finer auditory cues with computational models of the auditory periphery as ASR front-ends have also shown promise (e.g., Cohen, 1989[7], Jankowski *et al.*, 1995[8], Sandhu and Ghitza, 1995[9]). The success of these paradigms provides powerful motivation for exploring the integration of auditory cues into traditional signal processing approaches of speech representation.

The noise suppression abilities of the peripheral auditory system (PAS) has been demonstrated by many groups, both in physiological experiments (e.g., [10]) and in the responses of models of the PAS (e.g., [11]). This noise "robustness" is commonly attributed to the filtering action of the basilar membrane (BM). Linear time-invariant filtering approaches have allowed several groups [12, 13, 14] to obtain encouraging results for speech processing. Some researchers [15, 16] have incorporated BM nonlinearities into their models to accurately model data from physiological experiments. However, the lack of any compelling evidence that such an inclusion would help in applications, such as speech processing, provides little motivation for their integration into such models.

Recent work[17] in our group on the analysis of the effect of basilar membrane nonlinearities, in particular, the observed broadening of BM filtering with increasing stimuli level, indicates that BM nonlinearities act in a manner that strengthens vowel perception in noise but degrades consonant perception. This conflict suggests that improved robustness can be achieved by the use of dual filter-banks, broad bandwidth filters for robust representation of vocalic sounds and narrow bandwidth filters for robust representation of consonant-like sounds, in applications such as speech recognition. Our use of these dual filterbanks is the basis of dual-channel auditory modeling.

In this paper we formulate the dual-channel auditory spectrum (DCAS) model, propose a dual-channel auditory spectrum based cepstral coefficient (DCASCC) technique as well as provide the preliminary results obtained from various experiments conducted on our Entropic's HTK based ASR system.

2. THE DUAL-CHANNEL AUDITORY MODEL (DCAM)

The early auditory system constitutes the system that performs the initial transformation of auditory information (sound) from pressure differentials in air to the initial neural representations in the auditory nerve. These initial neural representations are then further processed in higher regions of the auditory nervous system.

Typically models of the early auditory system are composed of three cascaded stages: analysis, transduction, and reduction as shown in Figure 1.

Figure 1. Macroscopic view of a model of the early auditory system.

The analysis stage corresponds to transformation of sound into basilar membrane displacement. The varying stiffness of the BM results in in a filtering action that decomposes the

*This work was supervised by Prof. Amro El-Jaroudi, Associate Professor of Electrical Engineering, University of Pittsburgh, PA 15261. His name does not appear due to the ICASSP three paper limit.

sound into spectral bands akin to filterbank decomposition but differing in the non-linear intensity-dependent nature of BM filtering.

The transduction stage corresponds to the transduction of the cilia displacement into neural firings. The cilia displacement is initially transformed into variations of electrical potential which, in turn, modulate the release of neurotransmitter causing firings in the auditory nerves that populate the inner hair cells.

Finally these neural firings are passed on to the higher levels of the auditory nervous system, where the relevant information is extracted and processed. This forms the reduction stage of the model.

The auditory model from which the Dual-Channel Auditory Model is derived, originally proposed by Shamma *et al.* [18, 19], is shown in Figure 2. Briefly, it consists of a set of

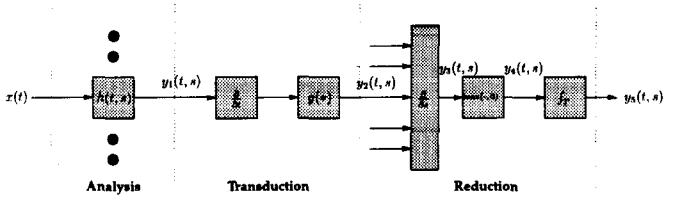

Figure 2. A model of the early auditory system.

filterbanks performing the initial frequency analysis, a time derivative and nonlinearity (we use a step function) corresponding to the transduction of inner hair cell cilia displacement, and finally a lateral inhibition network (LIN) reduction stage to model the information enhancement and extraction process. LINs are found in many biological systems and can be modeled (in their simplest form) as a single layer feedforward or recurrent neural network with mutually inhibitory weighting. In this model we use a two stage process: an initial spatial derivative stage that models the lateral coupling/inhibition amongst the LIN neurons and a half wave rectifier that models the neurons nonlinearity. This is followed by a temporal integration that mimics the inability of the auditory nervous system to follow rapid temporal variations.

Mathematically, if we denote the sound signal as $x(t)$, then the output of the analysis stage is given as:

$$y_1(t,s) = x(t) *_t h(t,s) \tag{1}$$

where $*_t$ represents convolution in time and $h(t,s)$ is the response of the cochlear filter at location s on the basilar membrane at any chosen sound level.

The output of the transduction stage can written as:

$$y_2(t,s) = g(\partial_t y_1(t,s)) \tag{2}$$

where $g(\cdot)$ represents the memoryless compressive step nonlinearity and $\partial_t \cdot$ the time derivative of the output of the analysis stage.

Finally, the LIN operation can be written as:

$$y_3(t,s) = \partial_s y_2(t,s) \tag{3}$$

$$= g'(\partial_t y_1(t,s))\partial_s \partial_t y_1(t,s) \tag{4}$$

$$y_4(t,s) = \max(y_3(t,s), 0) \tag{5}$$

$$y_5(t,s) = y_4(t,s) *_t \Pi(t) \tag{6}$$

where the final temporal integration stage has been denoted as a time convolution with a low-pass filter $\Pi(t)$.

The output of this stage $y_5(t,s)$, allowing for all our approximations and possible modeling errors, represents the sound as it would be portrayed to the higher regions of the auditory nervous system (ANS). This output is referred to as the *auditory spectrum*.

Our earlier work[17] analyzed the model of Figure 2 accounting for dynamic variations in the filtering of the filterbank first stage of the model. The results of this analysis suggested that vowel perception in noise improves as the BM filtering shows progressively broader tuning with increasing sound levels whereas consonant perception degrades with increasing sound levels.

This conflict indicates that improved robustness can be achieved by the use of dual filter-banks, broad bandwidth filters for robust representation of vocalic sounds and narrow bandwidth filters for robust representation of consonant-like sounds, in applications such as speech recognition.

Following this, the DCAM uses "broad" filters for voiced speech and "narrow" filters for unvoiced speech. The auditory model of Figure 2 is modified to allow for selection of the appropriate filter bank (narrow or broad) for the appropriate speech category. The resulting composite model is shown in Figure 3 as an expansion of the macroscopic model of Figure 1. Thus the resulting model is identical to the one

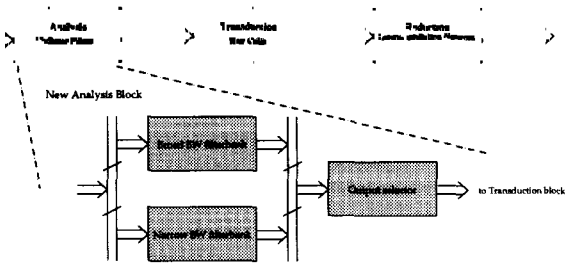

Figure 3. Composite model incorporating dual-channel processing.

shown in Figure 2 with the single filter bank being replaced by a new dual-channel filtering block as shown in Figure 4 and Figure 3.

We have developed a simple but effective hybrid channel selection algorithm that combines the Markel's SIFT unvoiced/voiced algorithm[20] with a threshold based detection algorithm on the auditory spectrum obtained from the broader filter bank, to provide a robust selection algorithm. The threshold based detection algorithm resulted from our empirical observation that, for progressively lower SNRs, the auditory spectrum output from the broader filter bank displayed increasing "flatness" for unvoiced input. This led us to devise a simple threshold algorithm based on the ratio of variance to mean of the auditory spectrum obtained from the broader filter bank. Figure 4 gives the flowchart for this algorithm.

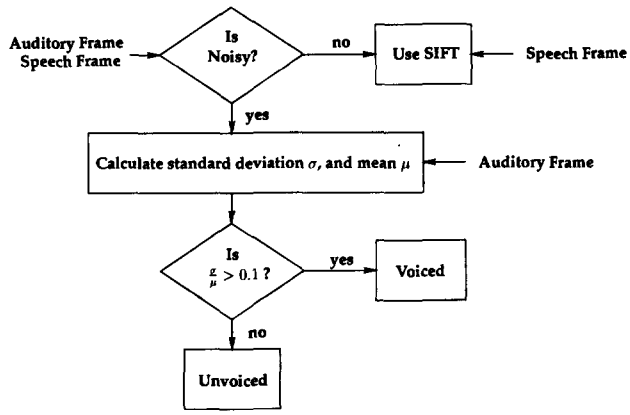

Figure 4. Flowchart for the hybrid voiced/unvoiced detector.

3. DCAS CEPSTRAL COEFFICIENTS

Previously we defined the output of the auditory model to be the "auditory spectrum" as it crudely represented the frequency "spectrum" seen by the ANS. Our features should, for this particular model, be related to and retain the information in the auditory spectrum. We choose an approach similar to that used for mel-frequency cepstral features, the essential difference being that our spectrum is not obtained as a Fourier transform of the input speech but as the output of an auditory model. The approach can be summarized as follows:

1. Obtain the auditory spectrum from the model.
2. Interpolate to obtain the appropriate frequency resolution.
3. Calculate the cepstral coefficients following Davis and Mermelstein (1980)[2].
4. Perform any needed post-processing such as cepstral weighting/liftering[21].

Our approach though unique follows similar work by Ghitza on his EIH model based ASR front end [13, 15] which used a LPC spectrum fitted to the EIH "spectrum" from which appropriate cepstral coefficients were determined. Another approach advocated by Seneff[12, 22] is the use of frequency content i.e., synchrony information as well as mean-rate information. These synchrony/mean-rate features were well suited to her model but would have been difficult to incorporate into our composite model. Again it should be emphasized that the choice of features is highly dependent on the auditory model and what it provides as its output. Our particular choice of features is at this point satisfactory to demonstrate the concept, we feel that further work is necessary to determine whether "better" features exist and how best to incorporate them.

4. EXPERIMENTS

Experiments were conducted on an HTK based phoneme recognizer. The vocabulary consisted of a reduced set of 48 phonemes taken from the TIMIT phoneme set. The data consisted of 1000 utterances arbitrarily chosen from the TIMIT corpus, 750 were used in the HMM training with 250 utterances used as an independent test set. No attempt was made to catalog the speaker characteristics.

For the experiments on noisy speech, utterances were individually processed by appropriate addition of pink Gaussian noise. The noise was created by initial generation of Gaussian white noise and subsequent passage of it through a second order low-pass filter with a nominal cutoff at 7200 Hz. The noise is suitably scaled before addition based on the desired SNR. The feature extraction process is identical to the clean speech case apart from this one additional "noise addition" preprocessing stage.

The experiment was carried out as follows:

1. Extract features from the 750 clean utterances.
2. Train the HMMs using these features.
3. Test the trained HMMs on the unseen 250 clean test utterances set.
4. Test the trained HMMs on the unseen 250 test utterances set after addition of noise.

For our experiments we used three different feature extraction schemes. Firstly, we used MFCC features with appended log-normalized energy as our baseline conventional ASR system. Our auditory model then furnished the DCASCC features also with appended log-normalized energy for our new ASR system. In an attempt to describe possible improvements over more mainstream single channel auditory models we disabled the selection algorithm and extracted features only from the broader filter bank channel, we will refer to such features as single channel auditory spectrum cepstral coefficients (SCASCCs). Log-normalized energy was also added to the SCASCCs. For the tests on noisy speech (the last stage) we choose SNR levels at 20dB, 10dB, 0dB and -5dB.

5. RESULTS

Table 1 shows the results for the experiments described above. As is clear, the robustness of the DCASCC features

	Clean	20dB	10dB	0dB	-5dB
MFCC	42.38%	19.11%	11.24%	6.05%	2.59%
SCASCC	28.33%	15.15%	9.96%	5.36%	3.51%
DCASCC	25.84%	14.38%	12.39%	8.34%	6.86%

Table 1. Preliminary results on our phoneme recognizer as %Correct

is remarkable. At -5dB SNR, the DCASCC front end outperforms the MFCC front-end by more than a factor of two and the SCASCC front-end by just under a factor of two. Much can be criticized about these results. Namely, that a more optimized system using more training data and the commonly used second order features might be a better baseline. In addition, 6.86% correct for DCASCC features at -5dB SNR is not acceptable for an ASR system. Moreover on clean speech the performance of auditory features is at best mediocre. While this criticism is valid, the fact remains that the concept of dual-channel auditory spectrum based features is workable and produces consistently better results than other known

techniques. Current efforts to further improve the performance of the proposed DCAM include the selection of more appropriate features from the auditory model and incorporating the model in a larger speech recognizer.

6. SUMMARY

In this paper we have brought together the theory introduced in our earlier work[17] to develop a workable means of incorporating dual-channel auditory spectrum based features into a state-of-art ASR system. We presented preliminary results which strongly support the validity and usefulness of the technique *vis-à-vis* both traditional mel-frequency cepstrum based features and single channel auditory modeling for the limited conditions of our experiments.

ACKNOWLEDGMENTS

We deeply appreciate the help extended to us by Li Deng, Steven Greenberg and Kuansan Wang by patiently answering our questions.

REFERENCES

[1] Alejandro Acero, *Acoustical and environmental robustness in automatic speech recogniton*, Kluwer Academic Publishers, Boston, 1993.

[2] Steven B. Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

[3] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[4] Hynek Hermansky and Nelson Morgan, "RASTA processing of speech", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, October 1994.

[5] Shoji Kajita and Fumitada Itakura, "Subband-autocorrelation analysis and its application for speech recognition", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1994, vol. 2, pp. 193–196.

[6] Shoji Kajita and Fumitada Itakura, "Robust speech feature extraction using SBCOR analysis", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1995, vol. 1, pp. 421–444.

[7] Jordan R. Cohen, "Application of an auditory model to speech recognition", *Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2623–2629, June 1989.

[8] Charles R. Jankowski Jr., Hoang-Doan H. Vo, and Richard Lippmann, "A comparision of signal processing front ends for automatic word recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 286–293, July 1995.

[9] Sumeet Sandhu and Oded Ghitza, "A comparitive study of mel-cepstra and EIH for phone recognition under adverse conditions", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1995, vol. 1, pp. 409–412.

[10] Murray. B. Sachs, H. F. Voigt, and Eric. D. Young, "Auditory nerve representation of vowels in background noise", *Journal of Neurophysiology*, vol. 50, no. 1, pp. 27–45, July 1983.

[11] Li Deng and C. Daniel Geisler, "A composite auditory model for processing speech sounds", *Journal of the Acoustical Society of America*, vol. 82, no. 6, pp. 2001–2012, December 1987.

[12] Stephanie Seneff, "A computational model for the peripheral auditory system: Application to speech recognition research", in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Tokyo, 1986, IEEE, pp. 1983–1986.

[13] Oded Ghitza, "Auditory representations as a basis for speech processing", in *Advances in Speech Signal Processing*, Sadaoki Furi and M. Mohan Sondhi, Eds., pp. 453–485. Markel Dekker, Inc., New York, 1992.

[14] Karen L. Payton, "Vowel processing by a model of the auditory periphery: A comparision to eighth-nerve responses", *Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 145–162, January 1988.

[15] Oded Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recogniton", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 115–132, January 1994.

[16] Christian Giguère and Phillip C. Woodland, "A computational model of the auditory periphery for speech and hearing research. I,II.", *Journal of the Acoustical Society of America*, vol. 95, no. 1, pp. 331–349, January 1994.

[17] Jayadev Billa and Amro El-Jaroudi, "An analysis of the effect of basilar membrane nonlinearities on noise suppression", *Journal of the Acoustical Society of America*, 1996, Under review. Draft available from the author by email request.

[18] Shihab A. Shamma, "Speech processing in the auditory system I,II", *Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1612–1632, November 1985.

[19] Kuansan Wang and Shihab Shamma, "Self-normalization and noise-robustness in early auditory representations", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 421–435, July 1994.

[20] J. D. Markel, "The sift algorithm for fundamental frequency estimation", *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 5, pp. 367–377, December 1972.

[21] Biing-Hwang Juang, Lawrence R. Rabiner, and Jay G. Wilpon, "On the use of bandpass liftering in speech recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 7, pp. 947–954, July 1987.

[22] Stephanie Seneff, "A joint synchrony/mean-rate model of auditory speech processing", *Journal of Phonetics*, vol. 16, pp. 55–76, 1988.