EARLY AUDITORY PROCESSING INSPIRED FEATURES FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Ozlem Kalinli and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Department of Electrical Engineering-Systems
University of Southern California
3750 McClintock Avenue, EEB 400, Los Angeles, California 90089.
kalinli@usc.edu, shri@sipi.usc.edu

ABSTRACT

In this paper, we derive bio-inspired features for automatic speech recognition based on the early processing stages in the human auditory system. The utility and robustness of the derived features are validated in a speech recognition task under a variety of noise conditions. First, we develop an auditory based feature by replacing the filterbank analysis stage of Mel-frequency cepstral coefficients (MFCC) feature extraction with an auditory model that consists of cochlear filtering, inner hair cell, and lateral inhibitory network stages. Then, we propose a new feature set that retains only the cochlear channel outputs that are more likely to fire the neurons in the central auditory system. This feature set is extracted by principal component analysis (PCA) of nonlinearly compressed early auditory spectrum. When evaluated in a connected digit recognition task using the Aurora 2.0 database, the proposed feature set has 40% and 18% average word error rate improvement relative to the MFCC and RelAtive SpecTrAl (RASTA) features, respectively.

1. INTRODUCTION

Hearing is one of the most highly developed senses in humans. The human auditory system can robustly localize, segment, and recognize sounds embedded in complex scenes. In contrast, machine recognition performance degrades drastically in various conditions such as in the presence of noise, speaker changes or overlapping sources. Despite years of intensive research in speech production and psychoacoustic analysis of human auditory system, the machine speech and audio processing methods still remain poor cousins to their biological counterparts. Understanding and modelling the information processing architectures in biological systems can offer the possibility of reducing the performance gap between human and machines in realistic conditions.

In the literature, there are signal representation methods based on physiological evidence such as linear predictive coding (LPC) and Mel-frequency cepstral coefficients (MFCC). While the LPC is related to the speech production model, the MFCC is based on a crude approximation of critical bands in the human auditory system. These features have been successfully used in speech recognition, audio classification, and auditory scene analysis, however, they are highly susceptible to noise. The perceptually inspired method called RelAtive SpecTrAl (RASTA) processing has been shown to improve robustness of speech recognition in the presence of noise [1]. It includes critical band analysis, temporal filtering, and equal loudness adjustment. It is designed to remove noise components by filtering out the slowly changing or steady-state factors interfering with the speech source.

There has also been research in the area of computational modelling of early and central stages of human auditory system for audio and speech processing. For example in [2, 3], it has been shown that their proposed early auditory model is robust to noise. This was used in [4] to extract MFCC-equivalent features by sampling the output of auditory spectrum at the channels corresponding to the MFCC's critical bands. In [5], robust processing was proposed by combining MFCC-type front end with an auditory based model. However, the speech recognition performance with the proposed feature set was not superior to MFCC [4, 5]. On the other hand, it has also been shown that multi scale spatio-temporal modulation features derived from central stages of auditory system are robust to noise in a classification task in [6], but these features are computationally very expensive for downstream processing since they produce a large dimensional tensor representation.

In this paper, we present biologically inspired robust speech processing algorithms based on human auditory system. As mentioned before, the multi-scale cortical representation of central auditory system is computationally very expensive [4, 6]. Hence we focus only on the early auditory (EA) processing which is computationally less expensive and has also been shown to be robust to noise. The contributions of this work are as follows. First, we develop an auditory processing based feature by replacing the triangular filter bank in MFCC feature extraction with a model that is more faithful to the processing stages in the EA system. The EA model used here consists of cochlear filtering, inner hair cell, and lateral inhibitory stages mimicking the process from basilar membrane to the cochlear nucleus in the auditory system. Then, a novel feature extraction algorithm is proposed which retains only the cochlear channel outputs that are more likely to fire neurons in the central auditory system by using principal component analysis (PCA). We also show empirically that an additional nonlinear compression modelling the outer hair cells has significant improvement on the speech recognition performance of the extracted feature set. The robustness of the developed features to a variety of noisy scenes is tested in a speech recognition task using the Aurora 2.0 database, and compared with state of the art MFCC and RASTA features. The experimental results show that the proposed feature set is more robust to noise compared to MFCC and RASTA features.

The paper is organized as follows. In Section 2, an overview of EA spectrum estimation along with the auditory

This research was supported by NSF, ONR and Army.

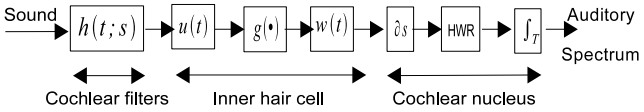

Figure 1: A computational model of processing in the early auditory system [2]

based feature extraction is provided. In Section 3, the experimental set-up together with the preliminary speech recognition results is discussed. Section 4 presents analysis of the auditory model, and explains the robust features obtained by post processing the EA model output. The experimental results are detailed in Section 5.

2. EARLY AUDITORY PROCESSING AND AUDITORY BASED FEATURES

In the human auditory system, when acoustic signal enters the ear, sound pressure waves create vibrations along the basilar membrane of cochlea. The cochlea separates the incoming signal frequencies by responding to different frequencies in different spatial locations along its length. Hence, the basilar membrane can be thought as a bank of band-pass filters $h(t;s)$ tonotopically ordered along the length of cochlea [2]. The spectral analysis performed by the cochlear filters is implemented as a bank of 128 overlapping constant-Q asymmetric band-pass filters [2]. The central frequencies of the band-pass filters are uniformly distributed along a logarithmic frequency axis (s).

The inner hair cell (IHC) stage transfers cochlear filter outputs into auditory nerve patterns. The IHC stage can be modelled in three steps: a high-pass filter $u(t)$ corresponding to fluid-cilia coupling, followed by a nonlinearity $g(\cdot)$ corresponding to ionic channel, and a low-pass filter $w(t)$ to model the leakiness of hair cell membrane [2]. Here, $g(\cdot)$ is implemented by a sigmoidal function, and $w(t)$ is implemented to represent phase-locking decrement in the auditory nerve beyond 2 kHz [2].

A hair cell fires when the potential builds up along the hair cell membrane. Auditory nerve fibers carry this neural spike to the cochlear nucleus of the central auditory system. In the cochlear nucleus, a lateral inhibitory network (LIN) detects discontinuities along the tonotopic axis [2].The LIN is modelled by a first-order spatial derivative (∂s) followed by a half wave rectifier (HWR) that models the nonlinearity of the neurons in the LIN. Here, the spatial derivative is approximated by a difference operation between adjacent frequency channels. The two-dimensional output, *auditory spectrum* [2], is obtained after leaky integration ($\int_T$) mimicking the inability of central neurons to follow rapid temporal changes.This stage is implemented as temporal filtering over a short time window, $\mu(t;\tau) = e^{-t/\tau}u(t)$, with time constant $\tau = 16$ ms. The block diagram of this early auditory processing is shown in Fig 1.

The widely used MFCC is based only on a crude approximation of basilar membrane filtering in the cochlea, and it has been shown empirically that it is highly susceptible to noise. Using a more accurate model of the auditory system can essentially help to obtain a better performance compared to the MFFC under noisy conditions. For this purpose, we introduce auditory based features (ABF) by replacing the triangular filter bank analysis stage used in MFCC computation

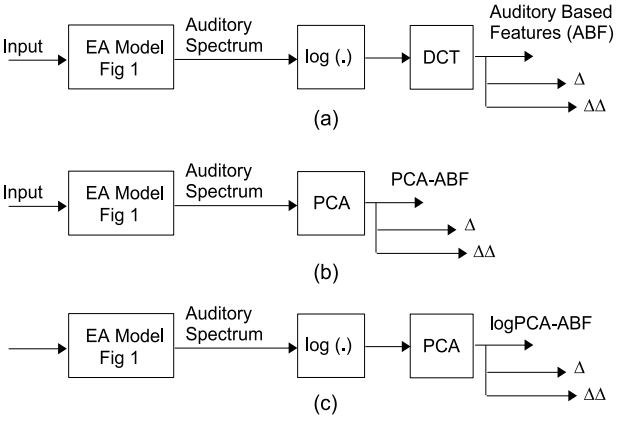

Figure 2: Block diagrams of feature extraction algorithms: (a) ABF feature extraction (b) PCA-ABF feature extraction (c) logPCA-ABF feature extraction

with aforementioned early auditory processing model. The ABF is expected to be robust to noise due to LIN and IHC stages in the EA model. The spatial derivative used in the LIN reduces the effect of noise due to the difference operation between adjacent channels, and the phase locking activity in IHC stage enhances the signal [3]. To obtain ABF, we compute the discrete cosine transform (DCT) of the logarithm of the *auditory spectrum*, and keep 13 of the coefficients as in the MFCC computation. The first (Δ) and second order time derivative ($\Delta\Delta$) features are appended to the raw features to form $39-$dimensional feature vector. In all of the proposed feature extraction methods in the following sections, Δ and $\Delta\Delta$ features are used together with raw features, unless stated otherwise. The block diagram of ABF extraction is summarized in Fig. 2(a), where "EA Model" box represents the early auditory process shown in Fig. 1.

3. EXPERIMENTAL SETUP AND PRELIMINARY RESULTS

To validate the new auditory based features, we perform speech recognition task on the Aurora 2.0 database [7] using the Hidden Markov Model Toolkit (HTK) [8]. The database consists of connected digits degraded with different noise conditions under different signal-to-noise ratios (SNR). We used 8440 clean utterances from 55 female and 55 male adults for training, and the recognition is done using the test sets with varying SNR levels (mis-matched training/testing). Training and testing follows the specifications detailed in [7].We created HMM word models for digits with 16 states per digit and 3 Gaussian mixtures per state. A three state silence model with 6 Gaussian mixtures per state and a one state short pause model which is tied to the middle stage of silence model are used. There are two sets of testing data; Set A and Set B. Set A contains the noise types of subway, babble, car and exhibition hall and Set B contains restaurant, street, airport , and train station noise at various SNR levels.

For MFCC feature extraction, we followed the specifications given in [7]. The 39-dimensional MFCC features consisting of 13 cepstral features plus Δ and $\Delta\Delta$ are used as a baseline. 23 channels were used during MFCC computation. The frame size was 25ms, and the frame shift was 10ms. The ABF extraction details were presented in Section 2.

The speech recognizer performance using both MFCC

and ABF features is shown in Fig 3. Here and in the preliminary results presented in Sec. 4.1 and 4.2, the data were degraded with subway noise for varying levels of SNR. We obtained similar results for other noise types as well (discussed later in Section 5 in detail). It can be observed from Fig 3 that replacing Mel-filterbank with a more accurate early auditory model improves the speech recognition performance under noisy conditions. These were our initial experiments to understand the potential of EA modelling, and to see the effect of using a more detailed feature model on speech recognition performance. The ABF treats all of the auditory channel outputs with equal importance. However, the channels with stronger stimulus might carry more information as explained in the next section. Thus, the auditory spectrum is post-processed before feeding it into the speech recognizer to further improve the noise robustness. The details are presented in the next section.

4. POST-PROCESSING OF AUDITORY SPECTRUM

4.1 Principal Components of Auditory Spectrum

The output of the early auditory model is transferred to the neurons in the central auditory system. The final stage of the early auditory model, leaky-integration, represents a simplified model of a leaky-integrate-and-fire (LIF) neuron model. These types of neurons accumulate the charges delivered by synaptic input, generate a spike when a threshold is reached, and reset the capacitive charge to zero after spike generation [9]. The stronger the stimulus, the higher is the chance of neuron getting fired.

The auditory spectrum obtained from the model presented in Fig. 1 represents the output of leaky integration. Here, it is assumed that the channel outputs that fire neurons carry the most significant information. Hence, we find the channel outputs that are more likely to generate a spike. Since a stronger stimulus has a better chance to generate a spike, the filter outputs are linearly transformed to a reduced dimension such that the reduced dimension features represent the strong components of the spectrum, which also means preserving the most of signal energy. To do this, we apply PCA [10] at the output of early auditory model. We retain only the most significant information by using PCA.

PCA is a dimension reduction technique that tries to obtain the best representation of the original data in the least squares sense in the projected space. Let $X = [x_1 x_2 \cdots x_N]$ be $d \times N$ data matrix, where $d = 128$ is the original data dimension and N is the sample size, and $W = [w_1 w_2 \cdots w_m]$ is the $d \times m$ transformation matrix, where $1 \leq m \leq d$. The goal of PCA is to find $\hat{W}$ such as:

$$\hat{W} = \arg\min \sum_{j=1}^{N} \|x_j - \sum_{i=1}^{m} (w_i^T x_j) w_i\|^2. \quad (1)$$

The problem reduces to finding eigenvalues of the sample covariance matrix $S = \frac{1}{N}\sum_{k=1}^{N}(x_k - u)(x_k - u)^T$, where u is the sample mean. The columns of $\hat{W}$ called principal components are the eigenvectors that correspond to the m largest eigenvalues of the data.

To set the number of principal components, we compute

$$\alpha_m = \frac{\sum_{k=1}^{m} \lambda_k^2}{\sum_{i=1}^{d} \lambda_i^2}. \quad (2)$$

α_m represents the portion of signal energy retained by keeping m principal components. We set m such that α_m is larger than 0.95, and we also consider the speech recognition performance.

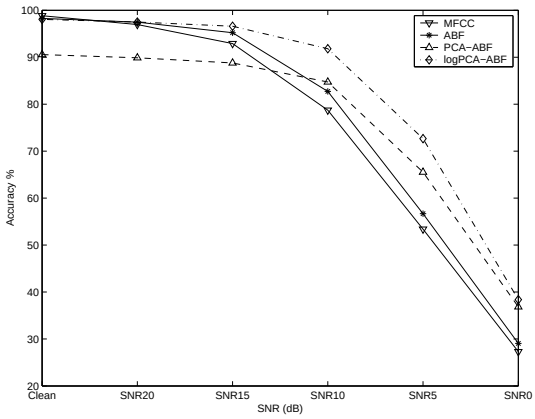

Figure 3: Speech recognition results with MFCC, ABF, PCA-ABF ($m = 10$), and logPCA-ABF ($m = 25$) for connected digits data with subway noise

The new feature extraction algorithm based on principal components of auditory spectrum, named as PCA-ABF, is summarized in Fig 2(b). The PCA transformation matrix W is learnt using the clean training data. The number of principal components retained is varied in the automatic speech recognition (ASR) experiments. The best ASR performance is achieved with $m = 10$ and $\alpha_{10} = 99\%$. In Fig 3, the speech recognition results with PCA-ABF are shown together with MFCC and ABF performance. The ASR results in Fig 3 show that PCA-ABF outperforms both ABF and MFCC for low SNR levels, whereas it performs poorer compared to ABF and MFCC for speech with moderate or high SNR levels. The PCA might be reducing the class discrimination for clean speech since only the significant channel outputs are retained, and this can cause some information loss about the source.However for speech with low SNR level, the gain achieved with the removal of noise components with PCA is higher than the source information loss, resulting in ASR performance improvement. Fig 3 shows that finding the principal components of EA model output is beneficial for low SNR levels. Thus, we kept PCA in our feature extraction algorithm but with an additional compression step modelling the outer hair cells (OHC) as explained in the next section.

4.2 Principal Components of Compressed Auditory Spectrum

The auditory nerves have limited dynamic range [11]. The dynamic range of basilar membrane and the neural response are compressed nonlinearly by the OHC. The OHC provide greater amplification to signals at low levels. We modified our model, and used logarithmic amplitude transformation to model the nonlinear compression due to OHC, and then applied PCA to the compressed auditory spectrum. This new feature set is called logPCA-ABF. The block diagram of logPCA-ABF feature extraction is shown in Fig. 2(c). The best ASR performance was achieved with $m = 25$ and $\alpha_{25} = 99\%$ for logPCA-ABF features. The ASR experiment results with logPCA-ABF features are shown in Fig. 3 together with the other features. Fig. 3 shows that with logPCA-ABF the performance degradation faced with using PCA-ABF feature set for speech with moderate or high SNR

Table 1: Speech Recognition Results with MFCC (Accuracy %)

	Set A					Set B				
	Subway	Babble	Car	Exhibition	Avg.	Restaurant	Street	Airport	Station	Avg.
Clean	98.83	98.97	98.81	99.14	98.94	98.83	98.97	98.81	99.14	98.94
SNR20	96.96	89.96	96.84	96.2	94.99	89.19	95.77	90.07	94.38	92.35
SNR15	92.91	73.43	89.53	91.85	86.93	74.39	88.27	76.89	83.62	80.79
SNR10	78.72	49.06	66.24	75.1	67.28	52.72	66.75	53.15	59.61	58.06
SNR5	53.39	27.03	33.49	43.51	39.36	29.57	38.15	30.69	29.74	32.04
SNR0	27.3	11.73	13.27	15.98	17.07	11.7	18.68	15.84	12.25	14.62
Avg.	74.69	58.36	66.36	70.30	67.43	59.4	67.77	60.91	63.12	62.80

Table 2: Speech Recognition Results with ABF (Accuracy %)

	Set A					Set B				
	Subway	Babble	Car	Exhibition	Avg.	Restaurant	Street	Airport	Station	Avg.
Clean	98.26	98.44	98.35	98.66	98.43	98.26	98.44	98.35	98.66	98.43
SNR20	97.47	91.56	96.91	96.1	95.51	90.45	96.19	90.9	95.44	93.25
SNR15	95.22	78.68	91.77	92.11	89.45	77.11	90.4	83.23	88.93	84.92
SNR10	82.72	57.3	70.86	76.68	71.89	58.26	69.1	60.87	66.39	63.66
SNR5	56.69	32.81	35.74	45.99	42.81	37.5	40.25	36.61	32.78	36.79
SNR0	29.06	15.46	15.78	18.87	19.79	16.55	21.29	19.48	13.14	17.62
Avg.	76.57	62.38	68.24	71.4	69.65	63.02	69.28	64.91	65.89	65.78
RI-M	7.43	9.65	5.59	3.7	6.82	8.92	4.69	10.23	7.51	8.01

level was resolved. Since the dynamic range is reduced due to the compression, we have to retain more principal components to have $\alpha_m = 0.99$. Thus, it can be expected that there is more detailed information compared to PCA-ABF method with increased number of principal components here, and this improves the results for clean speech. Also, with logPCA-ABF the ASR performance improved even more for speech with low SNR levels. The results with other noise types are discussed in Sec. 5.

5. EXPERIMENT RESULTS

The details of the speech recognition task were presented in Section 3. We used MFCC features as a baseline to compare our speech feature representations performance with. Also, we compared the speech recognition performance of our final feature set logPCA-ABF with RASTA features.

The speech recognition word accuracy results are given in Tables 1-4. For each noise type, we computed average word accuracy (denoted as "Avg.") in the tables with the results of all SNR levels including "clean" speech. We also computed relative word error rate (WER) improvement. In Table 2 and 4, "RI-M" and "RI-R" values show the relative WER improvement over MFCC and RASTA features, respectively.

The experiment results with MFCC feature set, and ABF set are given in Table 1 and Table 2, respectively. It can be observed that ABF performs better for noisy speech compared to MFCC. The average relative WER improvement obtained with ABF was 6.82% for Set A and 8.01% for Set B, resulting in overall 7.42% WER improvement over MFCC baseline. We believe that the slight performance degradation with ABF for clean speech is due to the lateral inhibitory network in the auditory model. In the LIN, while taking the difference of adjacent channels reduces the noise when the speech is noisy, this can cause information loss or introduce noise to clean speech. We can conclude from these experiments that using a better model of auditory system helps to improve speech recognizer performance when the speech is contaminated with noise.

The experiment results with RASTA and logPCA-ABF features are given in Table 3 and Table 4, respectively. The speech recognition performance of logPCA-ABF is compared with both MFCC and RASTA features. The average recognition result with Set A improved from 67.43% to 78.90%, and from 62.80% to 79.62% for Set B resulting in 35.2% and 45.2% relative WER improvement over the MFCC baseline. Similarly, with logPCA-ABF features the relative WER improvement over the RASTA feature performance was 19.94% and 17.47% for Sets A and B, respectively. It is clear that logPCA-ABF features work well for not only stationary noise types (i.e. car, exhibition hall [7]) but also non-stationary noise types (i.e. street, airport [7]). Overall, the logPCA-ABF features provide 40.2% and 18.71% relative WER improvement over the MFCC and RASTA features performance, respectively.

To compare all the results, we computed the average word accuracy over all noise types for each noise level condition, i.e. the average word accuracy for clean speech over all eight noise types. The results for all methods are presented in Fig 4. It is clear that the improvement gained with logPCA-ABF features is substantial, and it outperforms both MFCC and RASTA features in noisy conditions.

6. CONCLUSION AND FUTURE WORK

In this paper, we derived bio-inspired features for automatic speech recognition based on the processing stages in the early human auditory system. The derived features are validated in a speech recognition task in the presence of variety of noise types. First, we implemented an auditory based feature by replacing the Mel-filterbank analysis stage in MFCC feature extraction with an auditory model that consists of cochlear filtering, inner hair cell, and lateral inhibitory network stages. In our experiments, it was shown that the ABF was more robust to noise compared to MFCC. We derived a new set of features by post-processing the early auditory spectrum. In the experiments, it was shown that the selected features of nonlinearly compressed early auditory spectrum via PCA provided substantial improvement over both MFCC and RASTA features in noisy conditions. This is attributed to the noise suppressing feature of LIN, and signal enhance-

Table 3: Recognition Results with RASTA (Accuracy %)

	Set A					Set B				
	Subway	Babble	Car	Exhibition	Avg.	Restaurant	Street	Airport	Station	Avg.
Clean	98.7	98.93	98.99	99.1	98.93	98.7	98.93	98.99	99.1	98.93
SNR20	98.41	97.89	98.57	97.46	98.08	97.33	97.8	97.59	98.44	97.79
SNR15	96.68	93.85	95.14	95.17	95.21	94.4	94.03	94.96	94.96	94.59
SNR10	85.25	79.11	75.15	79.44	79.74	83.45	77.28	82.85	79.56	80.79
SNR5	56.14	49.01	38.2	43.61	46.74	57.48	48.01	54.48	46.52	51.62
SNR0	31.55	25.74	19.43	15.88	23.15	30.58	24.62	31.98	25.17	28.09
Avg.	77.79	74.09	70.91	71.78	73.64	76.99	73.45	76.81	73.96	75.3

Table 4: Recognition Results with logPCA-ABF (Accuracy %)

	Set A					Set B				
	Subway	Babble	Car	Exhibition	Avg.	Restaurant	Street	Airport	Station	Avg.
Clean	98.06	98.6	98.17	98.34	98.29	98.06	98.6	98.17	98.34	98.29
SNR20	97.5	96.46	97	96.27	96.81	95.17	96.68	94.37	96.17	95.6
SNR15	96.61	93.94	95.55	95.34	95.36	92.57	95.21	93.81	94.23	93.96
SNR10	90.81	86.24	85.89	88.09	87.76	87.66	86.73	87.08	86.08	86.89
SNR5	72.66	68.01	50.92	67.58	64.79	74	65.39	67.55	59.8	66.69
SNR0	38.37	32.9	22.49	27.68	30.36	44.45	31.08	39.33	30.23	36.27
Avg.	82.34	79.36	75	78.88	78.9	81.99	78.95	80.05	77.48	79.62
RI-M	30.21	50.43	25.69	28.9	35.2	55.63	34.68	48.97	38.92	45.2
RI-R	20.46	20.33	14.07	25.17	19.94	21.71	20.71	13.98	13.5	17.47

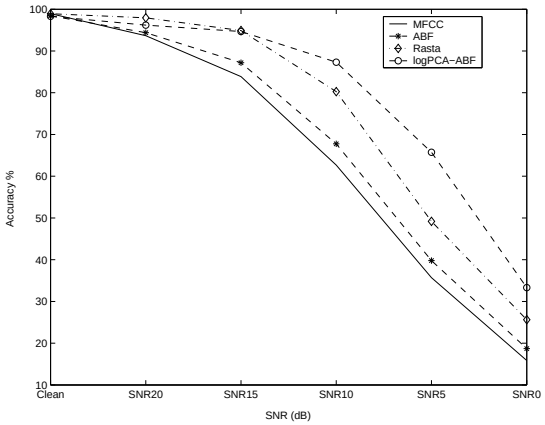

Figure 4: Performance comparison of all methods. Accuracy is the average of recognition results over all noise types. The proposed logPCA-ABF outperforms all other methods.

ment feature of IHC stages in the EA model. Also, by performing PCA on the nonlinearly compressed EA spectrum, only the channel outputs that are more likely to transmit information to the neurons in the central auditory system are selected, thereby removing insignificant channel outputs together with noise.

The experiment results showed the importance of two stages added to the early auditory model: *i*) the compression in the OHC *ii*) the selection of significant components of leaky integration taking place in the cochlear nucleus. As part of our future work, we plan to model the OHC compression more accurately as an adaptive model. We will also develop methods that can help us to code the spikes generated at the output of leaky integration such that it will represent relevant information more robustly.

REFERENCES

[1] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 578–589, October 1994.

[2] X. Yang, K. Wang, and S. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[3] K. Wang and S. Shamma, "Self-normalization and noise robustness in early auditory representations," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 3, pp. 421–435, 1994.

[4] W. Jeon and B.-H. Juang, "A study of auditory modeling and processing for speech signals," in *IEEE Int. Conf. Acoust., Speech Signal Proc.*, vol. 1, Pensylvania, USA, 2005, pp. 929–932.

[5] S. Ravindran, D. V. Anderson, and M. Slaney, "Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing," in *SAPA*, Pittsburgh, PA, September 2006.

[6] N. Mesgarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," in *IEEE Int. Conf. Acoust., Speech Signal Proc.*, Montreal, Canada, May 2004.

[7] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, Paris, France, September 2000.

[8] S. Young. (1989) Hidden markov model toolkit (HTK). [Online]. Available: http://htk.eng.cam.ac.uk/

[9] C. Koch, *Biophysics of Computation: Information Processing in Single Neurons*. New York, NY: Oxford University Press, 1999.

[10] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York, NY: Wiley-Interscience, 2001.

[11] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. New York, NY: John Wiley and Sons Inc., 2000.