ASR SYSTEMS IN NOISY ENVIRONEMENT: AUDITORY FEATURES BASED ON GAMMACHIRP FILTER USING THE AURORA DATABASE

Hajer Rahali, Zied Hajaiej, Noureddine Ellouze

Laboratoire des Systèmes et Traitement du Signal (LSTS)
Ecole Nationale d'Ingénieurs de Tunis, BP 37, Le Belvédère, 1002 Tunis, Tunisie
Hajer.Rahali@enit.rnu.tn
Zied.hajaiej@enit.rnu.tn
N.ellouze@enit.rnu.tn

ABSTRACT

This paper deals with the analysis of Automatic Speech Recognition (ASR) suitable for usage within noisy environment in various conditions. Recent research has shown that auditory features based on gammachirp filterbank (GF) are promising to improve robustness of ASR systems against noise. The behavior of parameterization techniques was analyzed from the viewpoint of robustness against noise. It was done for Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), Gammachirp Filterbank Cepstral Coefficient (GFCC) and Gammachirp Filterbank Perceptual Linear Prediction (GF-PLP). GFCC features have shown best recognition efficiency for clean as well as for noisy database. GFCC and GF-PLP features are calculated using Matlab and saved in HTK format. Training and testing for speech recognition is done using HTK. The above-mentioned techniques were tested with impulsive signals within AURORA databases.

Keywords: Gammachirp filter, Fourier transforms FFT, impulsive noise, MFCC, PLP.

1. INTRODUCTION

Automatic speech recognition systems are currently used in many applications in our everyday life. Due to the rapid development in this field all over the world we can see many systems and devices with voice input and output. Such a wide application area brings frequent usage of such systems also in noisy environment, so the issue of noise robustness represents the main topic of many research activities. An important drawback affecting most of the speech processing systems is the environmental noise and its harmful effect on the system performance. The presence of noise normally degrades the performance of speech recognition. A large amount of work has therefore been spent in this area and there exists a lot of technique that improves the speech recognition performances in noisy conditions. In the present work, we have applied an auditory filter that simulates the functions of the human auditory organ to process the impulse response and also some impulsive signals. The auditory models are generally a filterbank, none uniformly spaced in frequency and with non-uniform bandwidths, narrows at low frequencies, and broad at high frequencies, which converts the input speech signal into set of sub-band signals. The gammachirp auditory filter is an extension of the popular gammatone filter [7]; it has an additional frequency-modulation term to produce an asymmetric amplitude spectrum. The gammachirp has a much simpler impulse response than recent physiological models on cochlear mechanics [3]. Moreover, the chirp term in the gammachirp is consistent with physiological observations on frequency-modulations in measurements of the mechanical responses of the basilar membrane. Irino and Patterson have developed a theoretically optimal auditory filter, the gammachirp, whose parameters can be chosen to fit observed physiological and psychophysical data. The main parameters in the filterbank design are the frequency response, which defines the shape of the filters, the centre frequency and the bandwidth. These selected parameters based on the human auditory system. A more general approach is based on gammachirp filters, which involve dedicated design of mathematical forms of frequency response that match physiological experimental results. Similar to MFCC, this feature is usually referred to as Gammachirp frequency coefficient cepstra. The implementation of gammachirp filterbank shows consistent and significant performance gains in various noise types and levels. In this paper, we propose two techniques for parameterization speech signals based on a gammachirp filterbank following the approach used in the technical MFCC and PLP. For this we will develop a system for automatic recognition of isolated words with impulsive noise based on HMM\GMM. This work is concerned with the analysis and design of gammachirp filterbank and was successfully used for noise robust speech recognition. We propose a study of the performance of parameterization techniques GFCC and GF-PLP proposed in the presence of different impulsive noises. The sounds are added to the word with different signal-to-noise (20dB, 15dB, 10dB, 5 dB, 0 dB and -5 dB). The evaluation is done on the AURORA database. After extracting parameters we are interested to compare their performance with standard MFCC and PLP.

This paper is organized as follow; in the next section we briefly introduce the mel and gammachirp auditory filterbank. The processing steps of our gammachirp parameterization are described in section 4. Section 5 demonstrates simulations tested with new method. Finally, conclusions are given in section 6.

2. MFCC AND PLP FEATURES

As it is important for further discussion about noise robustness of studied features, the basic description of MFCC and PLP is presented, along with their principal block schemes in fig. 1 and 2.

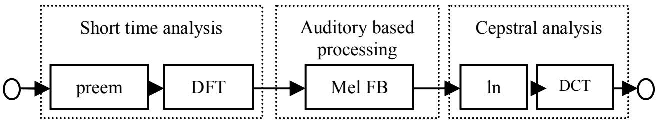

Fig. 1. Block scheme of MFCC feature extraction.

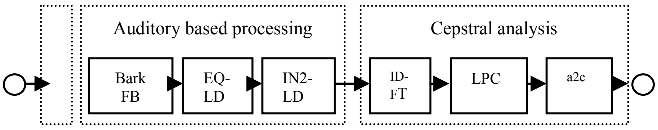

Fig. 2. Block scheme of PLP feature extraction.

Generally, both methods are based on three similar processing blocks: firstly, basic short-time Fourier analysis which is the same for both methods, secondly, auditory based filterbank (FB), and, thirdly, cepstral coefficients computation. Both methods use principally similar auditory modeling based on Mel- or Bark-scale with non-linear frequency warping bringing similar contribution to recognition results [12]. PLP uses also EQLD block (Equal LouDness) modifying the spectrum on the basis of frequency sensitivity of human hearing [19] and IN2LD block (INtensity-TO-LouDness) changing spectral dynamics according to the power-law of hearing [11]. MFCC changes frequency sensitivity only on the basis of standard pre-emphasis before STFT. The most significant difference between these two techniques lies in the final computation of cepstral coefficients. Autoregressive (AR) modeling is used in the case of PLP while MFCCs are computed directly using Discrete Cosine Transform (DCT) of the logarithmic auditory-based spectrum. The Mel Cepstral features are calculated by taking the cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. After pre-emphasizing the speech using a first order high pass filter and windowing the speech segments using a Hamming window of 20 ms length with 10 ms overlap, the DFT is taken of these segments. The magnitude of the Fourier Transform is then passed into a filterbank comprising of 25 triangular filters. The start and end points of these filters were calculated firstly by evenly spacing the triangular filters on the Mel-Scale and then using “eq. (1)" to convert these values back to the linear scale.

$$\mathrm{Mel}\ (f) = 2595 \,.\, \log_{10} (1 - \frac{f}{700}). \tag{1}$$

The Mel-filter bank is designed to simulate band pass filtering occurring in auditory system such that it is approximately linear up to 1 kHz and in actual frequency domain is logarithmic at higher frequencies. Such a model allows a constant bandwidth and constant spacing on the Mel-frequency scale and exploits the fact that the speech signal is stationary for short periods of time. It is modeled by constructing the required number of triangular band-pass filters with 50% overlap. Triangular band-pass filters are generated with Mel frequencies to be the centers of the triangles. The Bark filter bank is designed to simulate band pass filtering occurring in auditory system such that below 500 Hz the Bark scale becomes more and more linear.

3. THE GAMMACHIRP FILTER

The gammachirp filter is a good approximation to the frequency selective behavior of the cochlea [6]. It is an auditory filter which introduces an asymmetry and level dependent characteristics of the cochlear filters and it can be considered as a generalization and improvement of the gammatone filter. The gammachirp filter is defined in temporal domain by the real part of the complex function:

$$g_c(t) = a t^{n-1} e^{-2\pi B t}\, e^{j2\pi f_r t + jc\ln t + j\varphi}. \tag{2}$$

With

$$B = b.\,\mathrm{ERB}(f_r) = b.\,(24.7 + 0.108\ (f_r)). \tag{3}$$

Which is the equivalent rectangular bandwidth at frequency f_r , n is the filter order, f_r is the frequency modulation, a is some normalization constant, b is a parameter defining the envelope of the gamma distribution and c a parameter for the chirp rate.

3.1. Energy

The energy of the impulse response (t) is obtained with the following expression:

$$E_{n,B} = \|g_c\|^2 = < g_c, g_c > = a^2 \frac{\sigma(2n-1)}{4\pi B^{2n-1}}. \tag{4}$$

With σ(n) is the n-th order gamma distribution function. Thus, for energy normalization is obtained with the following expression:

$$A_{E_{n,B}} = \sqrt{\frac{4\pi B^{(2n-1)}}{\sigma(2n-1)}}. \tag{5}$$

3.2. Frequency response

The Fourier transform of the gammachirp in “eq. (2)” is derived as follows [3].

$$| \quad (f)| = \frac{a|\sigma(n)|}{\sigma(\,)} * \frac{\sigma(n)}{\left|2\pi\sqrt{((b\mathrm{ERB}(f_r))^2 + (f-f_r)^2}\right|^{n}}\, e^{c\theta}. \tag{6}$$

$$| \quad (f)| = | \quad | * \quad e^{c}. \tag{7}$$

$$\theta\ (f) = \arctan(\frac{f-}{b\mathrm{ERB}}). \tag{8}$$

| (f)| is the fourier magnitude spectrum of the gammatone filter, e^{c} is an asymmetric function since is anti-symmetric function centered at the asymptotic frequency. The spectral properties of the gammachirp will depend on the e^{c} factor; this factor has therefore been called the asymmetry factor. The degree of asymmetry depends on “c”. If “c” is negative, the transfer function, considered as a low pass filter, where c is positive it behave as a high-pass filter and if “c” zero, the transfer function, behave as a gammatone filter. In addition, this parameter is connected to the signal power by the expression, “eq. (9)” [2]:

$$c = 3.38 + 0.107\ \mathrm{Ps}. \tag{9}$$

3.3. Basic structure

Figure 3 shows a block diagram of the gammachirp filterbank. It is a cascade of three filterbanks: a gammatone filterbank, a lowpass-AC filterbank, and a highpass-AC filterbank [4].

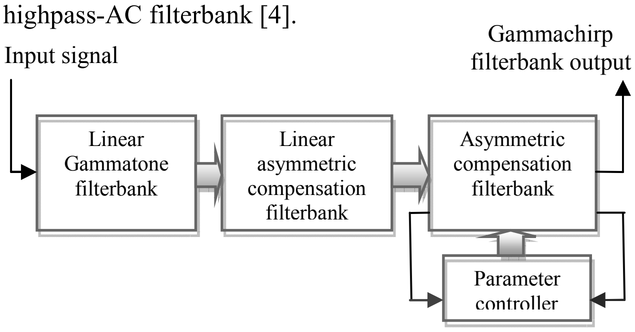
Input signal
Gammachirp filterbank output
Linear Gammatone filterbank
Linear asymmetric compensation filterbank
Asymmetric compensation filterbank
Parameter controller

Fig. 3. Structure of the Gammachirp filterbank.

The gammachirp filterbank consists of a gammatone filterbank and an asymmetric compensation filterbank controlled by a parameter controller with sound level estimation.

4. GFCC IMPLEMENTATION

With the gammachirp filterbank designed as described above, a frequency-time representation of the original signal, which is often referred to as a Cochleagram, can be obtained from the outputs of the filterbank. It is then straightforward to compute GFCC features from the Cochleagram. The remaining of this section presents the details of our GFCC implementation. Another popular feature set is the set of GF-PLP. Figure 4 shows the block diagram of the GFCC and GF-PLP.

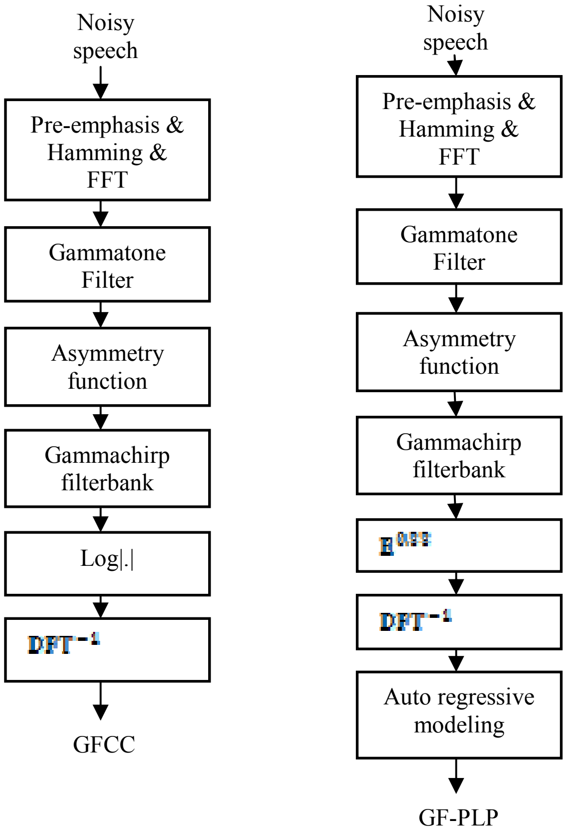

Fig. 4. Block diagram of the GF parameterization.

The GFCC are extracted from the speech signal according to the following steps; use the gammachirp filterbank defined in eq. (2) with 32 filters and the bandwidth multiplying factor F = 1.5 to bandpass the speech signal. After, estimate the logarithm of the short-time average of the energy operator for each one of the bandpass signals, and estimates the cepstrum coefficients using theDFT^{-1}. These steps are the main differences between MFCC and GFCC feature extraction.

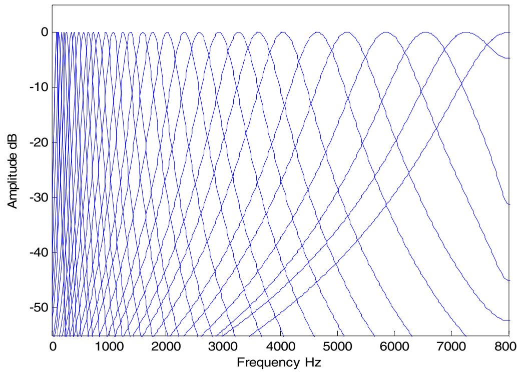

Fig. 5. A gammachirp filterbank with 32 filters.

The standard MFCC uses filters with frequency response that is triangular in shape (50% filter frequency response overlap). But, the proposed auditory GFCC use filters that are smoother and broader than the MFCC triangular filterbank (the bandwidth of the filter is controlled by the ERB curve and the bandwidth multiplication factor F). The main differences between the proposed filterbank and the typical one used for MFCC estimation are the type of filters used and their corresponding bandwidth. The gammachirp filterbank presented above, with bandwidths given by the ERB is a good approximation of the human auditory system. In this paper, we experiment with two parameters to create a family of gammachirp filterbanks:

- The number of filters in the filterbank.
- The bandwidth of the filters ERB (f). The bandwidth of the filter is obtained by multiplying the filter bandwidth curve ERB by the parameter F.

Experimental results provided in the next section show that both parameters are important for robust speech recognition. The range of parameters we have experimented is 20 – 40 for the number of filters and 1,0 – 2,0 for the bandwidth multiplying factor F. An example of the gammachirp filterbank employing 32 filters and with F = 1.5 is shown in fig. 5.

5. EXPERIMENT AND RESULT

In this section, we investigate the robustness of GFCC and GF-PLP in noise by artificially injecting various types of impulsive noise to the speech signal. We then present speech recognition experiments in noisy recording conditions.

5.1. AURORA task

AURORA is a noisy speech database, designed to evaluate the performance of speech recognition systems in noisy conditions. The AURORA task [18] has been defined by the European Telecommunications Standards Institute

(ETSI) as a cellular industry initiative to standardize a robust feature extraction technique for a distributed speech recognition framework. The initial ETSI task uses the TI-DIGITS database down sampled from the original sampling rate of 20 kHz to 8 kHz and normalized to the same amplitude level. Four different noises (Explosion, door slams, glass breaks and gunshots) have been artificially added to different portions of the database at signal-to-noise (SNR) ratios ranging from clean, 20dB to 0dB in decreasing steps of 5dB. The training set consists of 8440 different utterances split equally into 20 subsets of 422 utterances each. Each split has one of the four noises added at one of the six SNRs (20dB, 15dB, 10dB and 5dB, 0dB, -5dB). The test set consists of 4000 test files divided into four sets of 1000 files each.

5.2. Experimental setup

The analysis of speech signals is operated by using a gammachirp filterbank, in this work we use 32 gammachirp in each filterbank (of 4th order, n = 4), the filterbank is applied on the frequency band of [0 fs/2] Hz (where fs is the sampling frequency), after a pre-emphasis step and a segmentation of the speech signal into frames, and each frame is multiplied by a Hamming windows of 20ms. Each gammachirp filtering is obtained across two steps, in the first step, the speech frame is filtered by the correspondent 4th order gammatone filter, and in the second step we estimate the speech power and calculate the asymmetry parameter c. To evaluate the suggested techniques, we carried out a comparative study with different baseline parameterization techniques of MFCC and PLP implemented in HTK. The AURORA database is used for comparing the performances of the proposed feature extractor to the MFCC, PLP, GFCC and GF-PLP features, in the context of speech recognition. For the performance evaluation of our feature extractors, we have used the four noise of the AURORA corpus at seven different SNRs (clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB). The features extracted from clean and noisy database have been converted to HTK format using “VoiceBox” toolbox [19] for Matlab. Tables 1, 2, 3, 4 and 5 present the results from the series of recognition experiments to determine the effect of different noises on different features. In all the experiments, we use the constant characteristics, which are 5-states HMM with 9-Gaussian mixtures. We tested the performance in speech signal recognition with additive impulsive noise. This article represents an analysis of impulsive signals based on a gammachirp filterbank for the implementation of the words recognition. To evaluate the auditory gammachirp filterbank we have compared various techniques of standard parameterization (MFCC and PLP) with gammachirp parameterizations GFCC and GF-PLP, through recognition of word.

5.3. Results and discussion

The performance of the suggested parameterization methods GF-PLP and GFCC is tested on the AURORA databases using HTK. We use the percentage of word accuracy as a performance evaluation measure for comparing the recognition performances of the feature extractors considered in this paper. %: The percentage rate obtained. Tables 1, 2, 3, 4 and 5 present the average word accuracy (in %), averaged over all noise scenarios.

One Performance measures, the correct recognition rate (CORR) is adopted for comparison. They are defined as:

$$\% \text{ CORR} = \text{no. of correct labels/no. of total labels} * 100\%. \quad (10)$$

SNR		Clean			
	Features	MFCC	PLP	GF-PLP	GFCC
∞ dB		97.90	98.50	98.90	99.40

Table 1. Word accuracy (%) for clean speech.

SNR		Explosions			
	Features	MFCC	PLP	GF-PLP	GFCC
-5 dB		87.00	66.78	76.87	89.88
0 dB		80.79	78.98	77.67	90.56
5 dB		81.75	80.90	82.70	90.89
10 dB		83.60	81.37	85.68	91.05
15 dB		88.97	86.98	88.44	94.65
20 dB		93.80	92.30	97.98	98.67

Table 2. Word accuracy (%) for explosions.

SNR		Door slams			
	Features	MFCC	PLP	GF-PLP	GFCC
-5 dB		86.00	76.88	88.87	90.88
0 dB		86.80	78.38	89.27	91.73
5 dB		87.79	85.90	86.70	94.89
10 dB		88.75	86.77	87.22	94.95
15 dB		90.61	88.98	90.44	96.00
20 dB		96.97	92.66	97.98	98.61

Table 3. Word accuracy (%) for door slams.

SNR		Glass breaks			
	Features	MFCC	PLP	GF-PLP	GFCC
-5 dB		66.23	66.80	89.97	90.43
0 dB		75.32	89.38	90.27	91.43
5 dB		81.65	88.90	92.70	93.89
10 dB		98.90	91.37	93.42	94.45
15 dB		92.14	95.98	95.44	96.87
20 dB		96.76	96.98	98.23	98.21

Table 4. Word accuracy (%) for glass breaks.

SNR		Gunshots			
	Features	MFCC	PLP	GF-PLP	GFCC
-5 dB		76.23	66.34	87.97	90.43
0 dB		81.02	72.38	91.23	94.03
5 dB		82.45	81.90	90.76	95.84
10 dB		88.90	91.30	91.42	94.35
15 dB		93.14	95.22	92.44	97.87
20 dB		97.76	95.77	97.83	99.21

Table 5. Word accuracy (%) for gunshots.

In additive noise conditions the proposed method provides comparable results to that of the MFCC and PLP. In convolutive noise conditions, the proposed method provides consistently better word accuracy than all other methods. As we can see in the tables, the identification rate increases with speech quality, for higher SNR we have higher identification rate, the gammachirp filterbank based parameters are slightly more efficiencies than standard MFCC for noisy speech (98.67% vs 93.80% for 20 dB of SNR). Generally, we remark that with gammachirp filterbank the rates are slightly superior to mel triangular filterbank. In the previous section, we present the results of the gammachirp parameterization and the traditional methods MFCC, PLP. Between PLP and MFCC, MFCC performs slightly better than PLP in general. We can see the comparison between the two methods parameterization, these GFCC give better results in generalization and the better performance. The improvement is benefited from using a gammachirp filterbank instead of the triangular mel filterbank. In table 2 the recognition accuracy of the GFCC and GF-PLP, is 98.67% and 97.98%, respectively for 20 dB of SNR, but the results change the noise of another. From all the experiments, it was concluded that GFCC has shown best recognition performance compared to other feature extraction techniques because it incorporates gammachirp filter features extraction method. PLP features have also shown improvement in recognition performance as compared to MFCC. PLP features performed better from clean database because the signal was pre-emphasized by a simulated equal-loudness curve to match the frequency magnitude response of the ear as well as all signal components were perceptually equally weighted.

6. CONCLUSION

This paper reviewed the background and theory of the gammachirp auditory filter proposed by Irino and Patterson. The motivation for studying this auditory filter is to improve the signal processing strategies employed by automatic speech recognition systems. We have presented an approach of time-frequency analysis "auditory spectrogram" for speech. This takes account of characteristics of the ear. Were analyzed the impulsive noise based on a gammachirp filterbank, in word recognition. The gammachirp was compared to the Mel triangular filterbank. We observe that the worst results are those obtained with the basic modeling and best are those obtained with the model with gammachirp filterbank. Concerning this article, we presented the implementation of the gammachirp model of the cochlear filter. We validated this implementation by its use in analysis of some word with impulsive noise. The results gotten after application of this filter on the word show that this filter gives acceptable and sometimes better results by comparison at those gotten by other methods of parameterization such MFCC and PLP.

7. REFERENCES

[1] A. B. Poritz, "Hidden Markov models: A guided tour," *in Proc. Of the IEEE Int'l. Conf.* On Acoustics, Speech and Signal Processing (*ICASSP '88*), May 1988, pp. 7-13.

[2] T. Irino, and R. D. Patterson, "Temporal asymmetry in the auditory system," *J. Acoust. Soc. Am.* 99(4): 2316-2331, April, 1997.

[3] T. Irino, and R. D. Patterson, "A time-domain, Level-dependent auditory filter: The gammachirp," *J. Acoust.Soc. Am.* 101(1): 412-419, January, 1997.

[4] T. Irino, and M. Unoki, "An Analysis Auditory Filterbank Based on an IIR Implementation of the Gammachirp," *J. Acoust. SocJapan.* 20(6): 397-406, November, 1999.

[5] T. Irino, and R. D. Patterson, "A compressive gammachirp auditory filter for both physiological and psychophysical data," *J. Acoust Soc. Am.* 109(5): 2008-2022, may 2001.

[6] J. O. Smith III, and J.S. Abel, "Bark and ERB Bilinear Transforms," *IEEE Tran. On speech and Audio Processing*, Vol. 7, No. 6, November 1999.

[7] R. D. Patterson, and I. Nimmo-Smith, "Off-frequency listening and auditory-filter asymmetry," *J. Acoust. Soc. Am.*, Vol. 67, No. 1, pp. 229-245, 1980.

[8] B. R. Glasberg, and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, 47, 103-198, 1990.

[9] T. Irino, and M. Unoki, "A time-varying, analysis/synthesis auditory filterbank using the gammachirp," *IEEE Int. Conf. Acoust., Speech Signal Processing* (*ICASSP-98*), 3653-3656.

[10] University of Pennsylvania Linguistic Data Consortium. "Darpa-timit: a multi speaker's data base".

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech processing," *Proceedings of IEEE*, 77(2):257–286, 1989.

[12] J. W. Pitton, K. Wang, and B. H. Juang, "Time-frequency analysis and auditory modeling for automatic recognition of speech," *Proc. IEEE*, vol. 84, pp. 1199–1214, Sept. 1996.

[13] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 115–132, Jan. 1994.

[14] Skowronski M. D and Harris J. G, 2002, "Increased MFCC filter bandwidth for noise-robust phoneme recognition," *in Proc. ICASSP-02, Florida.*

[15] UMESH. S, COHEN. L, and NELSON. D, "Fitting the Mel scale," *In Proc. ICASSP,* 1999, vol. 1, p. 217-220.

[16] HERMANSKY. H, "Perceptual linear predictive (PLP) analysis of speech," *In Proc. JASA,* April 1990, vol. 87, no. 4.

[17] HÖNIG. F, STEMMER. G, HACKER. C, and BRUGNARA. F, "Revising Perceptual Linear Prediction (PLP)," *In Eurospeech* 2005, p. 2997-3000.

[18] H. G. Hirsch and D. Pearce, "The AURORA Experiment Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Condition," *ISCA ITRW ASR2000 Automatic Speech Recognition*: Challenges for the Next Millennium, France, 2000.

[19] M. Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB," Software, available [Mar. 2011] from, *www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.*