Auditory Model Based Optimization of MFCCs Improves Automatic Speech Recognition Performance

Saikat Chatterjee, Christos Koniaris and W. Bastiaan Kleijn

Sound and Image Processing Laboratory, School of Electrical Engineering
KTH - Royal Institute of Technology, Stockholm - 10044, Sweden
sach@kth.se, chris.koniaris@ee.kth.se, bastiaan.kleijn@ee.kth.se

Abstract

Using a spectral auditory model along with perturbation based analysis, we develop a new framework to optimize a set of features such that it emulates the behavior of the human auditory system. The optimization is carried out in an off-line manner based on the conjecture that the local geometries of the feature domain and the perceptual auditory domain should be similar. Using this principle, we modify and optimize the static mel frequency cepstral coefficients (MFCCs) without considering any feedback from the speech recognition system. We show that improved recognition performance is obtained for any environmental condition, clean as well as noisy.

Index Terms: MFCC, auditory model, ASR.

1. Introduction

An automatic speech recognition (ASR) system comprises two main tasks: feature extraction and pattern recognition. The feature extraction stage is designed to transform the incoming speech signal into a representation that serves as the input to a later pattern recognition stage. Feature extraction is a dimensionality reduction problem where the output representation should preserve the important aspects of the input speech signal relevant for speech recognition in any environmental condition, clean as well as noisy.

Different feature sets have been proposed in the literature, but the solutions remain ad hoc. We propose to define the features based on a perceptually relevant objective criterion. The human peripheral auditory system enhances the input speech signal for further processing by the central auditory system of the brain. Preprocessing of the input speech signal by the human auditory periphery forms a useful basis for designing an efficient feature set. Commonly used features use knowledge of the auditory system in an *ad hoc* manner. For example, several feature extraction methods perform auditory frequency filtering on a perceptually motivated frequency scale than a linear scale. Another example is the use of a logarithmic function to approximate the non-linear dynamic compression in the auditory system, which allows us to cover the large dynamic range between hearing threshold and uncomfortable loudness level. Using these two auditory motivated signal processing techniques, MFCCs were designed a few decades ago [1]. They are still universally used due to their computational simplicity as well as good performance. Importantly, the MFCCs do not use up-to-date quantitative knowledge of the auditory system.

Several attempts have been made to use quantitative auditory models in a practical ASR system processing chain [2]-[7]. In these techniques, the input speech signal is first processed through a readily available auditory model and then the output signal of the auditory model is formatted to use as an input to the pattern recognition stage of the ASR system. The direct use of an auditory model was shown to provide better speech recognition performance, but at the expense of higher computational complexity. In recent years, the research in quantitative modeling of the complex peripheral auditory system has reached a high level of sophistication [8]-[13], and it is appealing to use a sophisticated auditory model for designing efficient features. The feature set should not incur the higher computational complexity associated with a full auditory model.

In this paper, instead of the direct on-line use, we investigate the use of an auditory model to design improved MFCCs through off-line optimization. The optimized MFCCs are referred to as *modified MFCCs* (MMFCCs). The off-line approach helps to retain the computational simplicity of MMFCCs. Also, it avoids the difficulty of formatting the output of the auditory model for recognition. Comparing to traditional MFCCs, the MMFCCs have a similar structure as well as computational simplicity.

In our approach, the feature set is optimized in such a way that it emulates the behavior of the human auditory system. The implementation of our method relies on perturbation theory and does not consider any feedback from the ASR system. We conjecture that human-like classification of speech sounds is facilitated by similarity between the local geometries of two domains, the feature domain and the perceptual domain. For improved classification, the preservation of the data geometry near the class boundaries is most critical. This means that 'small' Euclidean distances must be similar in the two different domains, except for an overall scaling. The focus on small distances allows a complex perceptual distance to be reduced to a quadratic distance measure using a sensitivity matrix based analysis. The sensitivity matrix based analysis was first developed in the context of source coding [14]. In [15], the sensitivity matrix was used to simplify an auditory distance measure for audio coding. Here, we extend the sensitivity matrix paradigm to optimize a feature set. Using HTK, the optimized MMFCCs are shown to provide better recognition performance than traditional MFCCs for both clean and noisy acoustic conditions.

2. Maximizing Similarity between Spaces

Improvement in sound classification requires a feature representation that provides a good separation of sound classes in the feature space. Noting the high human recognition performance, it can be expected that the output of a sophisticated auditory model provides good separation of sound classes. Therefore, we optimize a feature set to better describe the inter-sound distances of a state-of-the-art auditory model. We conjecture that if the Euclidean distance between two acoustic features approximates the corresponding perceptual distortion for two different speech sounds, then the use of that acoustic feature generally leads to better classification in an ASR system. Ideally, this implies an isometry between the perceptual and feature domains. The mapping from the perceptual to feature domain would then be distance preserving.

2.1. Distance Preserving Measure

In practice, it is not be possible to design a feature set that leads to an accurate distance-preserving mapping from perceptual domain to feature domain. However, it is not required to preserve all the distances. For good classification, the preservation of the data geometry near the class boundaries is most critical. More generally, the preservation of small distances (reflecting the local geometry) near the classification boundary is important, whereas the preservation of large distances (reflecting the global geometry) is not required. In principle, to achieve better sound classification, we then simply desire to have the same small distances for the auditory domain and for the feature domain.

A feature set is a function of an input speech signal segment (or frame) and some adjustable design parameters. For example, to design MMFCCs, these design parameters can be the frequency warping parameter to change the shape of a filter bank (such as heights, widths, center frequency of filters), a parameter to change the shape of a compressing function (like logarithmic function), etc. The objective is to obtain a feature set with optimum parameters for which any small perturbation of the input speech signal segment leads to a Euclidean distance in the feature domain that best approximates the perceptual distortion indicated by the auditory model. Naturally this criterion has to hold for all speech segments. To measure the similarity of the auditory model distortion and the feature domain distance, a suitable objective measure needs to be designed that will provide a means of ensemble averaging over all speech segments and all perturbations. By optimizing the parameters, a higher similarity, through evaluating the objective measure, leads to a better feature set.

We now define an objective measure that relates between the perceptual and feature domains. Let us denote the signal vector for the j'th speech frame as $\mathbf{x}_j \in \mathbb{R}^N$, where $j \in J \subset \mathbb{Z}$, and the perceptual domain representation of $\mathbf{x}_j$ as $\mathbf{y} : \mathbb{R}^N \rightarrow \mathbb{R}^K$. We also denote the design parameters of a feature set by a vector $\mathbf{p} \in \mathbb{R}^S$. Then, we can denote the Q-dimensional feature derived from $\mathbf{x}_j$ using $\mathbf{p}$ as $\mathbf{c} : \mathbb{R}^N \times \mathbb{R}^S \rightarrow \mathbb{R}^Q$. The perceptual domain distortion is defined through a mapping as $\Upsilon : \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of non-negative reals. For the j'th speech frame, let us denote the l'th perturbed signal as $\hat{\mathbf{x}}_{j,l}$. Often the perceptual distortion measure is based on the L^2 norm of the difference between the perceptual domain signal $\mathbf{y}(\mathbf{x}_j)$ and its distorted version $\mathbf{y}(\hat{\mathbf{x}}_{j,l})$. In that case, $\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}) = \|\mathbf{y}(\mathbf{x}_j) - \mathbf{y}(\hat{\mathbf{x}}_{j,l})\|^2$. Using the L^2 norm, we can define a distance measure for the feature $\mathbf{c}(\mathbf{x}_j, \mathbf{p})$ as $\Gamma(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}, \mathbf{p}) = \|\mathbf{c}(\mathbf{x}_j, \mathbf{p}) - \mathbf{c}(\hat{\mathbf{x}}_{j,l}, \mathbf{p})\|^2$. Now, considering the finite sequence of speech frames $j \in J$ and a finite set of acoustic perturbations $l \in L_j$, the objective is to minimize a measure of dissimilarity between perceptual domain distortion and feature domain distortion with respect to the parameter set $\mathbf{p}$. To satisfy this objective, a suitable norm based measure can be defined as

$$\mathbf{O} = \sum_{j \in J} \sum_{l \in L_j} \left[\Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}) - \lambda\, \Gamma(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}, \mathbf{p})\right]^2, \tag{1}$$

where

$$\lambda = \frac{\sum_{j \in J} \sum_{l \in L_j} \Upsilon(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l})\; \Gamma(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}, \mathbf{p})}{\sum_{j \in J} \sum_{l \in L_j} \left(\Gamma(\mathbf{x}_j, \hat{\mathbf{x}}_{j,l}, \mathbf{p})\right)^2}. \tag{2}$$

Here λ is the necessary scaling to eliminate the effect of a scale mismatch between perceptual domain and feature domain. So, the objective is to minimize the norm based distance $\mathbf{O}$ with respect to the parameter vector $\mathbf{p}$.

2.2. Perturbation Analysis

While it is possible to minimize the objective measure of eq. (1) even for complex distortion measures, this can be computationally expensive. Since we are interested in small distances, we can approximate the perceptual and feature domain distortion measure using simpler quadratic measures, leading to a significant reduction in computational complexity and an increase in mathematical tractability. This approach is based on the sensitivity matrix framework [14], [15].

Let us omit the subscripts for notational brevity where no ambiguity exists. We assume that $\Upsilon(\mathbf{x}, \hat{\mathbf{x}})$ is analytic and $\Upsilon(\mathbf{x}, \mathbf{x}) = 0$. Then, for a sufficiently small perturbation $\hat{\mathbf{x}} - \mathbf{x}$, we can write

$$\Upsilon(\mathbf{x}, \hat{\mathbf{x}}) \approx \frac{1}{2}[\hat{\mathbf{x}} - \mathbf{x}]^T \mathbf{D}_\Upsilon(\mathbf{x})\, [\hat{\mathbf{x}} - \mathbf{x}], \tag{3}$$

where $\mathbf{D}_\Upsilon(\mathbf{x})$ is the sensitivity matrix whose elements are $\mathbf{D}_{\Upsilon,ij}(\mathbf{x}) = \left.\frac{\partial^2 \Upsilon(\mathbf{x}, \hat{\mathbf{x}})}{\partial \hat{x}_i\, \partial \hat{x}_j}\right|_{\hat{\mathbf{x}}=\mathbf{x}}$. In certain cases, such as the spectral auditory model of section 2.3, $\Upsilon(\mathbf{x}, \hat{\mathbf{x}})$ and $\mathbf{D}_\Upsilon(\mathbf{x})$ are known.

Next, we consider a simplification of the distortion in the feature domain i.e., $\Gamma(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{p})$. If the mapping $\mathbf{c}(\mathbf{x}, \mathbf{p})$ is analytic in $\mathbf{x}$, then we can use the Taylor series expansion to make a local approximation around $\mathbf{x}$ as

$$\mathbf{c}(\hat{\mathbf{x}}, \mathbf{p}) = \mathbf{c}(\mathbf{x}, \mathbf{p}) + \mathbf{A}(\mathbf{p})\; [\hat{\mathbf{x}} - \mathbf{x}], \tag{4}$$

where $\mathbf{A}(\mathbf{p})$ is a $Q \times N$-dimensional matrix as $\mathbf{A}(\mathbf{p}) = \left.\frac{\partial \mathbf{c}(\mathbf{x}, \mathbf{p})}{\partial \hat{\mathbf{x}}}\right|_{\hat{\mathbf{x}}=\mathbf{x}}$. We can then write the distortion in the feature domain as

$$\begin{aligned} \Gamma(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{p}) &= \|\mathbf{c}(\mathbf{x}, \mathbf{p}) - \mathbf{c}(\hat{\mathbf{x}}, \mathbf{p})\|^2 \\ &= [\hat{\mathbf{x}} - \mathbf{x}]^T \mathbf{A}(\mathbf{p})^T\, \mathbf{A}(\mathbf{p})\, [\hat{\mathbf{x}} - \mathbf{x}]. \end{aligned} \tag{5}$$

2.3. A Spectral Auditory Model

In this paper, we optimize the MMFCCs to minimize the norm based measure of eq. (1). The MMFCCs are designed using the power spectrum of the input speech signal. Therefore, for optimization, we use the spectral auditory model developed by van de Par, et al. [13] which is referred to as the van de Par auditory model (VAM). The VAM is a psycho-acoustic masking model that accounts for simultaneous processing of sound signals with different frequencies. To use the VAM, we consider the input signal $\mathbf{x}$ as the power spectrum of a speech frame. The VAM consists of several frequency channels, in each of which the ratio of distortion power to masker power is calculated. Then, the ratios of all the frequency channels are combined together to account for the spectral integration property of the human auditory system. Let $\mathbf{H}$ be a diagonal N-dimensional matrix whose diagonal is formed by the frequency response of the outer and middle ear filter. In the same fashion, a diagonal $\mathbf{G}_i$ is defined, so that the frequency response of the i'th channel Gamma-tone auditory filter forms its diagonal. For the VAM, the diagonal sensitivity matrix is

$$\mathbf{D}_\Upsilon(\mathbf{x}) \approx 2\frac{C_s L_e}{N} \sum_i \frac{[\mathbf{G}_i\, \mathbf{H}]^T [\mathbf{G}_i\, \mathbf{H}]}{\frac{1}{N}[\mathbf{G}_i\, \mathbf{H}\, \mathbf{x}]^T [\mathbf{G}_i\, \mathbf{H}\, \mathbf{x}] + C_a}, \tag{6}$$

where C_s and C_a are constants calibrated based on measurement data, and L_e is a constant to account for the influence of temporal integration time in the human auditory system on frame duration.

It is important to mention that each speech frame is independently analyzed in the VAM. Therefore, the use of VAM is appropriate for optimizing a static feature. Note that due to the inability to model the auditory response across speech frames, the use of the VAM is inappropriate for optimizing the temporal dynamic features, such as velocity and acceleration. However, it is possible to compute the dynamic features from any static feature using standard regression method.

3. Modified MFCCs

We first generalize the definition of the MFCCs to render a set of features with adjustable parameters $\mathbf{p}$. We refer to this new set of features as *modified MFCCs* (MMFCCs). Let the N-dimensional vector $\mathbf{x} = [x_0\ x_1, \cdots, x_n, \cdots, x_{N-1}]^T$ be the power spectrum of a Hamming windowed speech frame. Then the steps of evaluating the MMFCCs are as follows:

1. Calculation of the energy in each channel:

$$\begin{aligned} z_m &= \mathbf{x}^T\ \mathbf{w}_m(\alpha) \\ &= \sum_{n=0}^{N-1} x_n \times w_{m,n}(\alpha)\,, 0 \le m \le M-1, \end{aligned} \tag{7}$$

where $\mathbf{w}_m(\alpha)$ is the N-dimensional vector denoting the triangular filter of the m'th channel and satisfies $\sum_{n=0}^{N-1} w_{m,n}(\alpha) = 1$. M is the total number of channels with a typical value of $M = 26$. The shape of a triangular filter depends on the extent of frequency warping. The warped frequency scale [16] is given as

$$f_{warp} = 2595 \times \log_{10}\left(1 + (f/\alpha)\right), \tag{8}$$

where α is the warping factor and f is the frequency in Hz. An increase in α leads to a decrease in the extent of warping. For the MMFCCs, α is a parameter to optimize to achieve better recognition performance. In the case of MFCCs, the triangular filters are designed using the *mel* frequency scale where $\alpha = 700$ [16].

2. Compression of the dynamic range of the energy in each channel:

$$s_m = \log_{10}\left[\sum_{r=1}^{R} b_r\,(z_m)^r\right],\ 0 \le m \le M-1, \tag{9}$$

where $\sum_{r=1}^{R} b_r = 1$ and $b_r \ge 0$. For the MMFCCs, we optimize the polynomial coefficients $\{b_r\}_{r=1}^{R}$. In the case of MFCCs, $R = 1$ and $b_1 = 1$ [1]. We note that eq. (9) implies that our results are scale dependent and require proper normalization.

3. De-correlation using the DCT to evaluate Q-dimensional MMFCC feature vector:

$$c_q = \sum_{m=0}^{M-1} s_m \times \cos\left[q\,(m+0.5)\,\frac{\pi}{M}\right], 1 \le q \le Q. \tag{10}$$

A typical value of feature vector dimension is $Q = 12$.

3.1. Optimization of the MMFCCs

The parameters that we optimize to obtain the MMFCCs are $\mathbf{p} = \left[\alpha, \{b_r\}_{r=1}^{R}\right]$. To optimize the parameters, we need to minimize the objective measure $\mathbf{O}$ of eq. (1). This objective measure is a function of the sensitivity matrix based perceptual domain distortion of eq. (3) and the feature domain distortion of eq. (5). To evaluate the perceptual domain distortion, we need a closed form sensitivity matrix $\mathbf{D}_\Upsilon(\mathbf{x})$ which is given by the VAM as shown in eq. (6). We also need a closed form $\mathbf{A}(\mathbf{p})$ for evaluating the feature domain distortion. For an MMFCC feature, the elements of the matrix $\mathbf{A}(\mathbf{p})$ are

$$\begin{aligned} A_{qn} &= \frac{\partial c_q}{\partial x_n} = \frac{\partial c_q}{\partial s_m}\,\frac{\partial s_m}{\partial z_m}\,\frac{\partial z_m}{\partial s_n} \\ &= \sum_{m=0}^{M-1} \cos\left[q\,(m+0.5)\,\tfrac{\pi}{M}\right] \\ &\quad \times \frac{\sum_{r=1}^{R} r\,b_r\,(z_m)^{r-1}}{\ln 10\ \times \sum_{r=1}^{R} b_r\,(z_m)^r}\, w_{m,n}(\alpha). \end{aligned} \tag{11}$$

Table 1: Phone recognition accuracy (in %) of static 12-dimensional MFCC and MMFCC features using TIMIT

Feature	Number of Gaussian mixtures/state							
	1	2	4	6	8	10	12	14
MFCC	43.19	47.00	48.63	49.57	50.38	51.04	51.60	51.96
MMFCC	45.13	48.63	50.16	51.28	52.10	52.63	52.92	53.30

It is interesting to jointly optimize all the parameters through a closed-form/iterative solution, such as using gradient descent search technique. This requires a closed form gradient expression $\frac{d\mathbf{O}}{d\mathbf{p}}$, which is not easy to evaluate due to the intricate relationship existing between the measure of $\mathbf{O}$ and the parameter vector $\mathbf{p} = \left[\alpha, \{b_r\}_{r=1}^{R}\right]$. Therefore, we use a simple increment-based linear search technique and optimize the parameters one by one. We first optimize $\{b_r\}_{r=1}^{R}$ and then α. For both the cases of wide-band (sampling frequency 16 kHz) and narrow-band (sampling frequency 8 kHz) speech, we use a 32 ms Hamming windowed speech frame with 10 ms frame shift. To evaluate the MMFCCs, we use $M = 26$ and $Q = 12$. The power spectrum of each frame is computed using a standard DFT based periodogram technique and the power spectrum is perturbed with i.i.d Gaussian noise at different SNRs ranging from 120 to 130 dB. Using an increment-based linear search, we evaluate the minimum value of the measure of eq. (1) and find that a polynomial order of $R = 2$ is sufficient; the values of the polynomial coefficients are $b_1 = 0.1$ and $b_2 = 0.9$. Next we search for the optimum α. For wide-band speech and narrow-band speech, we find the optimum values are $\alpha = 900$ and $\alpha = 1100$, respectively. We note that standard MFCCs use $b_1 = 1$ and $\alpha = 700$ irrespective of the sampling frequency of input speech, choice of the window length and shift, and the feature dimension (Q) and number of channels (M) [1], [16].

4. Recognition Results

Using the HTK toolkit, we performed phone and word recognition experiments to compare between MFCC and MMFCC features. The static 12-dimensional MFCC feature set was extracted using the same setup as that used to extract the 12-dimensional MMFCC feature set. Using the standard approach, 39-dimensional feature vectors were evaluated. To the static features, we appended the log energy of a speech frame and the velocity and acceleration of the features.

We first compared the performance of 12-dimensional static features through a clean speech phoneme recognition experiment. In this case, we used the TIMIT database where the speech is sampled at 16 kHz. HTK training and testing were performed using the training set and the test set of TIMIT respectively. The TIMIT transcriptions are based on 61 phones. Following convention, the 61 phones were folded onto 39 phones as described in [17]. To train the HMMs, we used three states per phoneme and the performance is shown in Table 1 for a varying number of Gaussian mixtures per state. We used Gaussian mixtures with diagonal covariance matrices. From Table 1, it can be noted that MMFCCs outperform MFCCs for any number of mixtures. In case of the 39-dimensional feature vectors, the performance improvement of using MMFCCs over MFCCs was always positive, but small for clean speech phone recognition.

Next we considered the 39-dimensional feature vectors for robust word and phone recognition experiments where clean speech training and noisy speech testing were performed. For the robust

Table 2: Robust word recognition accuracy (in %) of 39-dimensional MFCC and MMFCC features using Aurora 2

Feature	Test Set a				Test Set b				Test Set c	
	set 1	set 2	set 3	set 4	set 1	set 2	set 3	set 4	set 1	set 2
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Train-Station	Subway	Street
	SNR = 20 dB									
MFCC	95.46	96.67	96.12	94.88	96.87	96.28	96.78	96.33	93.28	94.41
MMFCC	96.49	97.49	97.08	96.42	97.73	97.04	97.58	97.04	94.96	95.62
	SNR = 10 dB									
MFCC	85.05	86.49	83.39	81.70	87.69	83.89	87.50	84.45	74.88	76.00
MMFCC	87.07	88.51	87.00	85.25	87.81	86.06	87.77	87.84	79.49	78.36

Table 3: Robust phone recognition accuracy (in %) of 39-dimensional MFCC and MMFCC features at 10 dB SNR

Feature	Performance in Accuracy				
	Clean	White	Pink	Babble	Volvo
MFCC	68.11	37.03	40.51	46.25	59.71
MMFCC	68.34	43.65	46.67	48.94	61.91

recognition, we used cepstrum mean and variance normalization (CMVN) on the feature sets [18]. For the robust word recognition experiment, we used the Aurora 2 database where the speech is sampled at 8 kHz and the sub-datasets of test set are corrupted by different noise types at varying SNRs. The standard configuration of the HTK setup was used where HMMs were trained using 16 states per word and three Gaussian mixtures per state (diagonal covariance matrices). The robust word recognition performance for 39-dimensional MFCCs and MMFCCs are shown in Table 2 at the testing conditions of 20 dB and 10 dB SNRs. We note that MMFCCs perform better than MFCCs for all the sub-datasets corrupted with different noises. In the case of clean speech word recognition, the improvement of 39-dimensional MMFCCs over MFCCs was small like in the case of clean speech phone recognition.

Finally, we consider a robust phone recognition experiment where the clean test speech database of TIMIT was corrupted with additive noise. We used the following noise types from the NoiseX-92 database: white, pink, babble and car (volvo) noise. The test speech database was corrupted by adding each noise at 10 dB SNR. The HMMs consisted of three states per phoneme and 20 Gaussian mixtures per state. The robust phone recognition performance for MFCCs and MMFCCs are shown in Table 3 and we note that MMFCCs perform better than MFCCs for all noise types.

5. Conclusions

Our development of MMFCCs shows that the use of a sophisticated auditory model can lead to a simple feature set that provides improved speech recognition performance for any environmental condition. The success of our perceptual-distance preserving measure in optimizing features suggests that the auditory system provides as output a signal representation that is 'efficient' for speech recognition. As we developed the static MMFCCs using a static spectral auditory model, further investigation should consider the optimization of dynamic features using a spectro-temporal auditory model, such as that presented in [11].

6. Acknowledgments

This work is supported by EU FP6 ACORNS project. The authors wish to thank V. Surayanarayana. K of Indian Institute of Science, Bangalore, India.

7. References

[1] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 28, No. 4, pp. 357-366, Aug. 1980.

[2] J.R. Cohen, "Application of an auditory model to speech recognition," *J. Acoust. Soc. Amer.*, pp. 2623-2629, Vol. 85 (6), June 1989.

[3] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech, Audio Proc.*, vol. 2, No. 1, pp. 115-132, Jan 1994.

[4] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech, Audio Proc.*, vol. 5, No. 5, pp. 451-464, Sept. 1997.

[5] D.S. Kim, S.Y. Lee and R.M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech, Audio Proc.*, vol. 7, No. 1, pp. 55-69, Jan 1999.

[6] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Amer.*, pp. 2040-2050, Vol. 106 (4), Oct. 1999.

[7] M. Holmberg, D. Gelbart and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing," *IEEE Trans. Speech, Audio Proc.*, vol. 14, No. 1, pp. 43-49, Jan 2006.

[8] S. Seneff, "A joint synchrony/mean-rate model of auditory processing," *J. Phonet.*, pp. 55-76, Vol. 85 (1), Jan 1988.

[9] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Acoust. Soc. Amer.*, pp. 702-711, Vol. 79 (3), March 1988.

[10] J.M. Kates, "Two-tone suppression in a cochlear model," *IEEE Trans. Speech, Audio Proc.*, vol. 3, No. 5, pp. 396-406, Sept. 1995.

[11] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Amer.*, pp. 3615-3622, Vol. 99 (6), Jun 1996.

[12] A.J. Oxenham, "Forward masking: Adaptation or integration?," *J. Acoust. Soc. Amer.*, pp. 732-741, Vol. 109 (2), Feb 2001.

[13] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen and S.H. Jensen, "A Perceptual model for sinusoidal audio coding based on spectral integration" *EURASIP J. Applied Signal Proc.*, vol. 9, pp. 1292-1304, 2005.

[14] W.R. Gardner and B.D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters," *IEEE Trans. Speech and Audio Proc.*, vol. 3, No.5, pp. 367-381, Sept 1995.

[15] J.H. Plasberg and W.B. Kleijn, "The sensitivity matrix: using advanced auditory models in speech and audio processing," *IEEE Trans. Audio, Speech, Language Proc.*, vol. 15, No. 1, pp. 310-319, Jan 2007.

[16] J.W. Picone, "Signal modeling techniques in speech recognition," *Proc. IEEE*, pp. 1215-1247, Vol. 81, No. 9, Sept. 1993.

[17] K.F. Lee and H.W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 37, No. 11, pp. 1641-1648, Nov. 1989.

[18] J. Droppo and A. Acero, "Environmental robustness," *Handbook of Speech Processing*, Springer, pp. 658-659, Oct. 2007.