IDIAP RESEARCH REPORT

Mel-Cepstrum Modulation Spectrum (MCMS) Features for Robust ASR

Vivek Tyagi [a] Iain McCowan [a]
Hervé Bourlard [a,b] Hemant Misra [a,b]

IDIAP–RR 03-47

September 2003

to appear in
Proceedings of IEEE ASRU 2003, St. Thomas, US Virgin Islands

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP, Martigny, Switzerland
[b] EPFL, Lausanne, Switzerland

IDIAP Research Report 03-47

Mel-Cepstrum Modulation Spectrum (MCMS) Features for Robust ASR

Vivek Tyagi Iain McCowan Hervé Bourlard Hemant Misra

September 2003

to appear in
Proceedings of IEEE ASRU 2003, St. Thomas, US Virgin Islands

Abstract. In this paper, we present new dynamic features derived from the modulation spectrum of the cepstral trajectories of the speech signal. Cepstral trajectories are projected over the basis of sines and cosines yielding the cepstral modulation frequency response of the speech signal. We show that the different sines and cosines basis vectors select different modulation frequencies, whereas, the frequency responses of the delta and the double delta filters are only centered over 15Hz. Therefore, projecting cepstral trajectories over the basis of sines and cosines yield a more complementary and discriminative range of features. In this work, the cepstrum reconstructed from the lower cepstral modulation frequency components is used as the static feature. In experiments, it is shown that, as well as providing an improvement in clean conditions, these new dynamic features yield a significant increase in the speech recognition performance in various noise conditions when compared directly to the standard temporal derivative features and C-JRASTA PLP features.

1 Introduction

A central result from the study of the human speech perception is the importance of slow changes in speech spectrum for speech intelligibility [14]. A second key to human speech recognition is the integration of phonetic information over relatively long intervals of time. Speech is a dynamic acoustic signal with many sources of variation. As noted by Furui [7, 8], spectral changes are a major cue in phonetic discrimination. Moreover, in the presence of acoustic interference, the temporal characteristics of speech appear to be less variable than the static characteristics [2]. Therefore, representations and recognition algorithms that better use the information based on the specific temporal properties of speech should be more noise robust [5, 6]. Temporal derivative features [7, 8] of static spectral features like filter-bank, Linear Prediction (LP) [10] , or mel-frequency cepstrum [11] have yielded significant improvements in ASR performances. Similarly, the RASTA processing [5] and cepstral mean normalization (CMN) techniques, which perform cepstral filtering, have provided a remarkable amount of noise robustness.

Using these temporal processing ideas, we have developed a speech representation which factorizes the spectral changes over time into slow and fast moving orthogonal components. Any DFT coefficient of a speech frame, considered as a function of frame index with the discrete frequency fixed, can be interpreted as the output of a linear time-invariant filter with a narrow-bandpass frequency response. Therefore, taking a second DFT of a given spectral band, across frame index, with discrete frequency fixed, will capture the spectral changes in that band with different rates. This effectively extracts the modulation frequency response of the spectral band.

The use of term “modulation” in this paper is slightly different from that used by others [2, 12]. For example, “modulation spectrum” [2] uses low-pass filters on time trajectory of the spectrum to remove fast moving components. In this work, we instead apply several band-pass filters in the mel-cepstrum domain. In the rest of this paper, we refer to this representation as the Mel-Cepstrum Modulation Spectrum (MCMS).

Our work is reminiscent of the cepstral time matrices in [3, 4]. However, in our work, we start from the modulation spectrum of the speech signal and show that it can be seen as linear transformation of the DFT outputs of the cepstral trajectories. As it is well known that the cepstral parameters are highly uncorrelated, the proposed cepstral modulation frequency based features performed better than the modulation frequency based features. Using lower cepstral modulation frequency components, we can reconstruct the cepstrum for each frame and use it as static feature. In this work, we propose using the MCMS coefficients as dynamic features for robust speech recognition. Comparing the proposed MCMS features to standard delta and acceleration features, it is shown that while both implement a form of band-pass filtering in the cepstral modulation frequency, the bank of filters used in MCMS have better selectivity and yield more complementary features. In [1], we have used MCMS coefficients in the (10-15) Hz cepstral modulation frequency range. In this work we achieve further improvement in the recognition accuracies by using MCMS coefficients in the range (3-22) Hz.

In Section 2, we first give an overview and visualisation of the modulation frequency response. The visual representation is shown to be very stable in presence of additive noise. The proposed reconstructed cepstrum and MCMS dynamic features are then derived in Section 3. Finally, Section 4 compares the performance of the MCMS features with standard temporal derivative features in recognition experiments on the Numbers database for non-stationary noisy environments.

2 Modulation Frequency Response of Speech

Let $X[n,k]$ be the DFT of a speech signal $x[m]$, windowed by a sequence $w[m]$. Then, by rearrangement of terms, the DFT operation could be expressed as,

$$X[n,\ k] = x[n] * h_k[n] \tag{1}$$

where $'*'$ denotes convolution and,

$$h_k[n] = w[-n]e^{\frac{j2\pi kn}{M}} \tag{2}$$

From (1) and (2), we can make the well-known observation that the k^{th} DFT coefficient $X[n,\ k]$, as a function of frame index n, and with discrete frequency k fixed, can be interpreted as the output of a linear time invariant filter with impulse response $h_k[n]$. Taking a second DFT, of the time sequence of the k^{th} DFT coefficient, will factorize the spectral dynamics of the k^{th} DFT coefficient into slow and fast moving modulation frequencies. We call the resulting second DFT the "Modulation Frequency Response" of the k^{th} DFT coefficient. Let us define a sequence $y_k[n] = X[n,\ k]$. Then taking a second DFT of this sequence over P points, gives

$$Y_k(q) = \sum_{p=0}^{P-1} y_k(n+p)e^{\frac{-j2\pi qp}{P}}\ ,\ \ q \in [0,\ P-1] \tag{3}$$

$$Y_k(q) = \sum_{p=0}^{P-1} X[n+p,\ k]e^{\frac{-j2\pi qp}{P}}$$

where $Y_k(q)$ is termed the q^{th} modulation frequency coefficient of k^{th} primary DFT coefficient. Lower $q's$ correspond to slower spectral changes and higher $q's$ correspond to faster spectral changes. For example, if the spectrum $X[n,\ k]$ varies a lot around the frequency k, then $Y_k(q)$ will be large for higher values of modulation frequency, q. This representation should be noise robust, as the temporal characteristics of speech appear to be less variable than the static characteristics. We note that $Y_k(q)$ has dimensions of $[T^{-2}]$.

To illustrate the modulation frequency response, in the following we derive a modulation spectrum based on (3), and plot it as a series of modulation spectrograms. This representation emphasizes the temporal structure of the speech and displays the fast and slow modulations of the spectrum. Our modulation spectrum is a four-dimensional quantity with time n (1), linear frequency k (1) and modulation frequency q (3) being the three variables.

Let $C[n,\ l]$ be the real cepstrum of the DFT $X[n,\ k]$.

$$C[n,\ l] = \frac{1}{K-1}\sum_{k=0}^{K} \log(\mid X[n,\ k] \mid)e^{\frac{+j2\pi kl}{K}}\ ,\ \ l \in [0,\ K-1] \tag{4}$$

Using a rectangular low quefrency lifter which retains only the first 12 cepstral coefficients, we obtain a smoothed estimate of the spectrum, noted $S[n,\ k]$.

$$\log S[n,\ k] = C[n,\ 0] + \sum_{l=1}^{L-1} 2C[n,\ l]\cos(\frac{2\pi lk}{K}) \tag{5}$$

where we have used the fact that $C[n, l]$ is a real symmetric sequence. The resulting smoothed spectrum $S[n,\ k]$ is also real and symmetric. $S[n,\ k]$ is divided into B linearly spaced frequency bands and the average energy, $E[n,\ b]$, in each band is computed.

$$E[n,\ b] = \frac{1}{K/B}\sum_{i=0}^{K/B-1} S[n,\ b\frac{K}{B}+i]\ ,\ \ b \in [0,\ K/B-1] \tag{6}$$

Let $M[n,\ b,\ q]$ be the magnitude modulation spectrum of band b computed over P points.

$$\begin{gathered} M[n,\ b,\ q] = \mid \textstyle\sum_{p=0}^{P-1} E[n+p,\ b]e^{\frac{-j2\pi pq}{P}} \mid\ , \\ with\ q \in [0, P],\ b \in [0,\ K/B-1] \end{gathered} \tag{7}$$

The modulation spectrum $M[n,\ b,\ q]$ is a 4-dimensional quantity. Keeping the frequency band number b fixed, it can be plotted as a conventional spectrogram. Figures 1 and 2 show conventional spectrograms of a clean speech utterance and its noisy version at SNR6. Whereas, the Figures 3 and 4 show modulation spectrograms of the same clean and noisy utterance as above. The stability of modulation spectrogram towards additive noise can be easily noticed in these figures. The figures consists of 16 modulation spectrograms, corresponding to each of 16 frequency bands in (6), stacked on top of each other. In our implementation, we have used a frame shift of 3ms and the primary DFT window of length 32ms. The secondary DFT window has a length $P = 41$ which is equal to 3ms*40=120ms. This size was chosen, assuming that this would capture phone specific modulations rather than average speech like modulations. We divided $[0,\ 4kHz]$ into 16 bands for the computation of modulation spectrum in (7). For the second DFT the Nyquist frequency is 333.33 Hz. We have only retained the modulation frequency response up to 50 Hz as there was negligible energy present in the band [50Hz, 166Hz]. For every band, we have shown the modulation spectrum with $q \in [1,\ 20]$, which corresponds to the modulation frequency range, [0Hz, 160Hz]

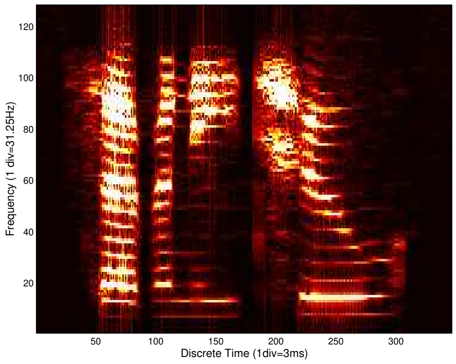

Figure 1: *Conventional Spectrogram of a clean speech utterance.*

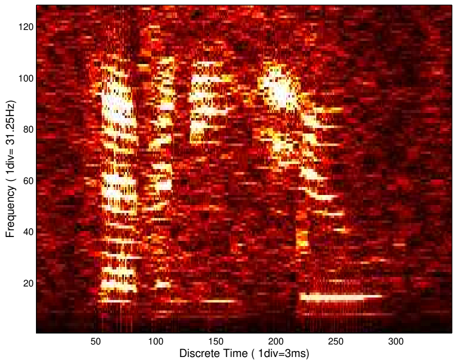

Figure 2: *Conventional Spectrogram of a noisy speech utterance at SNR6.*

3 Mel-Cepstrum Modulation Spectrum Features

As the spectral energies $E[n, b]$ in adjacent bands in (6) are highly correlated, the use of the magnitude modulation spectrum $M[n, b, q]$ as features for ASR would not be expected to work well (this has been verified experimentally). Instead, we here compute the modulation spectrum in the cepstral domain, which is known to be highly uncorrelated. The resulting features are referred to here as Mel-Cepstrum Modulation Spectrum (MCMS) features.

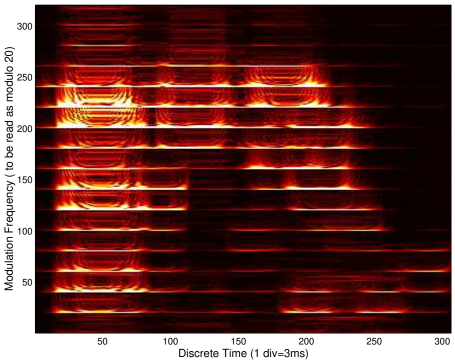

Figure 3: *Modulation Spectrum across 16 bands for a clean speech utterance. The above figure is equivalent to 16 modulation spectrums corresponding to each of 16 bands. To see q^{th} modulation frequency sample of b^{th} band, go to number $(b-1)*6+q$ on the modulation frequency axis.*

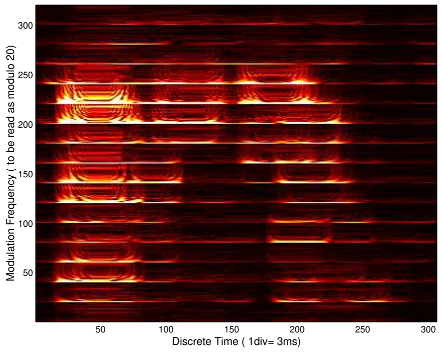

Figure 4: *Modulation Spectrum across 16 bands for a noisy speech utterance at SNR6. The above figure is equivalent to 16 modulation spectrums corresponding to each of 16 bands. To see q^{th} modulation frequency sample of b^{th} band, go to number $(b-1)*6+q$ on the modulation frequency axis.*

Consider the modulation spectrum of the cepstrally smoothed power spectrum $\log(S[n,\,k])$ in (5). Taking the DFT of $\log(S[n,\,k])$ over P points and considering the q^{th} coefficient $M^{'}[n,k,q]$, we obtain,

$$M^{'}[n,\,k,\,q] = \sum_{p=0}^{P-1} \log(S[n+p,\,k])e^{\frac{-j2\pi pq}{P}} \tag{8}$$

Using (5), (8) can be expressed as,

$$\begin{aligned} M^{'}[n,\,k,\,q] &= \textstyle\sum_{p=0}^{P-1} C[n,\,0]e^{\frac{-j2\pi pq}{P}} \\ &+ \textstyle\sum_{l=1}^{L-1}\cos(\frac{2\pi kl}{K})\underbrace{\displaystyle\sum_{p=0}^{P-1} 2C[n+p,\,l]e^{\frac{-j2\pi pq}{P}}} \end{aligned} \tag{9}$$

In (9) we identify that the under-braced term is the cepstrum modulation spectrum. Therefore, $M^{'}[n,\,k,\,q]$ is a linear transformation of the cepstrum modulation spectrum. As cepstral coefficients are mutually uncorrelated, we expect the cepstrum modulation spectrum to perform better than the power spectrum modulation spectrum $M^{'}[n,\,k,\,q]$. Therefore, we define:

$$MCMS_{DFT}[n,\ k,\ q] = \sum_{p=0}^{P-1} C[n+p,\ l]e^{\frac{-j2\pi pq}{P}} \tag{10}$$

An alternative interpretation of the MCMS features, is as filtering of the cepstral trajectory in the cepstral modulation frequency domain. Temporal derivatives of the cepstral trajectory can also be viewed as performing filtering operation. Figure 5 shows the cepstral modulation frequency response of the filters corresponding to first and second order derivatives of the MFCC features, while Figure 6 shows few of the filters employed in the computation of the MCMS features. On direct comparison, we notice that both of the temporal derivative filters emphasize the same cepstral modulation frequency components around 15Hz. This is in contrast to the MCMS features, which emphasize different cepstral modulation frequency components. This further illustrates the fact that the different MCMS features carry complementary information.

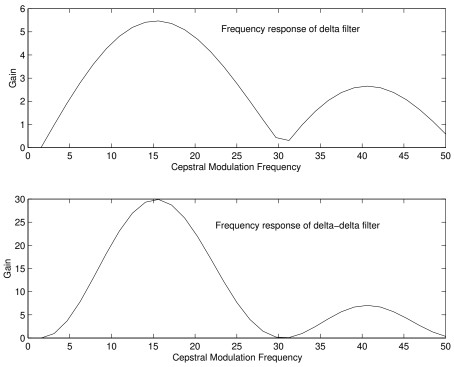

Figure 5: *Cepstral Modulation Frequency responses of the filters used in computation of derivative and acceleration of MFCC features*

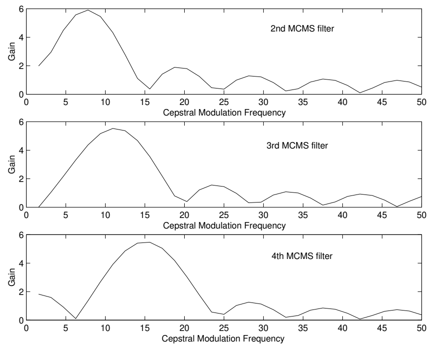

Figure 6: *Cepstral Modulation Frequency responses of the filters used in computation of MCMS features*

Let $MCMS_{DCT}[n,\ k,\ q]$ be the q^{th} DCT coefficient of the k^{th} cepstral trajectory taken across P frames.

$$MCMS_{DCT}[n,\ k,\ q] = \sum_{p=0}^{P-1} C[n+p-P/2,\ k]cos^{\frac{j\pi(q)(p+0.5)}{P}} \tag{11}$$

$$with\ q \in [0, P-1],\ k \in [0,\ L-1]$$

In our experiments, we noticed that the higher MCMS coefficients usually degraded the speech recognition performance. Therefore, using first $P^{'}$ MCMS coefficients, where $P^{'} < P$, we removed the high cepstral modulation frequency components from the raw cepstrum to obtain a relatively smoother cepstral trajectories, $C_{reconstructed}[n,\, k]$.

$$C_{reconstructed}[n,\, k] = \frac{1}{P}MCMS_{DCT}[n,\, k,\, 0]$$

$$+\sum_{q=1}^{P'-1}\frac{2}{P}MCMS_{DCT}[n,\, k,\, q]cos\frac{j\pi(P/2)(q+0.5)}{P} \tag{12}$$

$$with\, k \in [0,\, L-1]$$

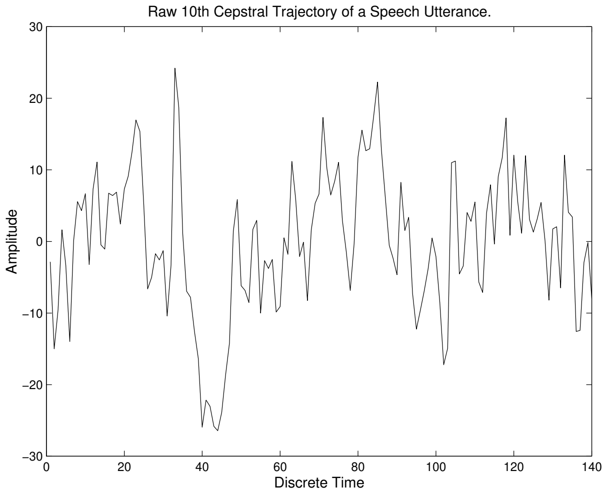

Figure 7: *Trajectory of 10th Cepstral coefficient.*

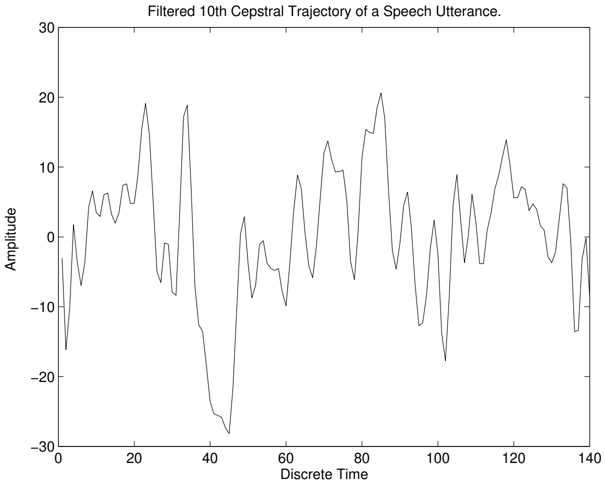

Figure 8: *Trajectory of the reconstructed 10th Cepstral coefficient.*

In Figure 7, we show the trajectory of the 10^{th} cepstral coefficient of a clean speech utterance. In figure 8, the corresponding reconstructed trajectory is shown. Due to the smoothness of the trajectory, we can notice the absence of high cepstral modulation frequency components in it. This is desirable as these components usually degrade the speech recognition performance. In the following experiments, the reconstructed smoothed cepstrum $C_{reconstructed}[n, k]$ has been used as the static feature in place of raw cepstrum.

4 Recognition Experiments

In order to assess the effectiveness of the proposed MCMS features for speech recognition, experiments were conducted on the Numbers corpus. Four feature sets were generated :

MFCC+Deltas: 39 element feature vector consisting of 13 MFCCs (including 0^{th} cepstral coefficient) with cepstral mean subtraction and variance normalization and their standard delta and acceleration features.

PLP + Deltas C-JRASTA Processed 39 element feature vector consisting of 13 PLPs which have been filtered by constant J-RASTA filter and their standard delta and acceleration features.

$MCMS_{DFT}$**:** 78 element feature vector consisting of first three real and imaginary $MCMS_{DFT}$ coefficients derived form a basis of sines and cosines as in (10) with variance normalization

MFCC+MCMS: 78 element feature vector consisting of 13 MFCCs (including 0^{th} cepstral coefficient) which have been reconstructed from lower MCMS coefficients as in (12) with their first five $MCMS_{DCT}$ dynamic features as in (11) with variance normalization.

We note that a direct comparison between MCMS, MFCC and PLP features of the same dimension (39 in each case) was presented in [1]. In this work we investigate the use of a greater range of MCMS filters covering (3-22) Hz of cepstral modulation frequency (6 MCMS filters).

The speech recognition systems were trained using HTK on the clean training set from the original Numbers corpus. The system consisted of 80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state. The context length for the MCMS features has been kept at 11 frames which corresponds to 120ms. The MCMS features used in the experiment cover a cepstral modulation frequency range from 3Hz to 22 Hz. This range was selected as it yielded the best recognition performance. The MCMS features used in these experiments are computed over a 120ms long time window. To verify the robustness of the features to noise, the clean test utterances were corrupted using Factory and Lynx noises from the Noisex92 database [13]. The results for the baseline and MCMS systems in various levels of noise are given in Tables 1 and 2.

From the results in Tables 1 and 2 we see a number of interesting points. First, the MCMS features lead to a significant decrease in word error for clean conditions. This is an important result, as most noise robust features generally lead to some degradation in clean conditions (such as RASTA-PLP, for example). Moreover, these features show greater robustness in moderate to high levels of non-stationary noise than both the MFCC and RASTA-PLP features, which are common features in state-of-the-art robust speech systems.

Table 1: *Word error rate results for factory noise*

SNR	MFCC+Deltas	C-JRASTA PLP	$MCMS_{DFT}$	MFCC+$MCMS_{DCT}$
Clean	7.2	9.4	5.2	5.0
12 dB	16.7	13.4	10.3	10.1
6 dB	26.2	25.0	18.4	19.3

Table 2: *Word error rate results for lynx noise*

SNR	MFCC+Deltas	C-JRASTA PLP	$MCMS_{DFT}$	MFCC+$MCMS_{DCT}$
Clean	7.2	9.4	5.2	5.0
12 dB	15.8	11.2	7.6	7.4
6 dB	20.5	17.6	11.2	11.4

5 Conclusion

In this paper we have proposed a new feature representation that exploits the temporal structure of speech, which we referred to here as the Mel-Cepstrum Modulation Spectrum (MCMS). These features can be seen as the outputs of an array of band-pass filters applied in the cepstral modulation frequency domain, and as such factor the spectral dynamics into orthogonal components moving at different rates. In our experiments, we found that a context length of 120ms for the computation of MCMS features, performs the best. This is in agreement with the findings of Hermansky [5] and Milner [3, 4], where they integrate spectral information over longer periods of time. In these experiments we have used 6 MCMS coefficients which cover the cepstral modulation frequency in the range (3, 22) Hz. In experiments, the proposed MCMS dynamic features are compared to standard delta and acceleration temporal derivative features and constant J-RASTA features. Recognition results demonstrate that the MCMS features lead to significant performance improvement in non-stationary noise, while importantly also achieving improved performance in clean conditions.

6 Acknowledgements

The authors would like to thank Hynek Hermansky for his insightful comments on this work. The authors also wish to thank DARPA for supporting this work through the EARS (Effective, Affordable, Reusable Speech-to-Text) project.

References

[1] Vivek Tyagi, Iain Mccowan, Hervé Bourlard, Hemant Misra, “ On factorizing spectral dynamics for robust speech recognition,” to appear in Proc. Eurospeech, Geneva, Switzerland 2003.

[2] B.E.D. Kingsbury, N. Morgan and S. Greenberg, “ Robust speech recognition using the modulation spectrogram,” Speech Communication, vol. 25, Nos. 1-3, August 1998.

[3] B.P. Milner and S.V. Vaseghi, “ An analysis of cepstral time feature matrices for noise and channel robust speech recognition”, Proc. Eurospeech, pp. 519-522, 1995.

[4] B.P. Milner, “Inclusion of temporal information into features for speech recognition”, Proc. ICSLP, PP. 256-259, 1996.

[5] H. Hermansky and N. Morgan, “RASTA Processing of Speech,” IEEE Trans. on Speech and Audio Processing, 2: 578-589, October, 1994.

[6] Chin-Hui Lee, F.K. Soong and K.K. Paliwal, eds. “Automatic Speech and Speaker Recognition”, Massachusetts, Kluwer Academic, c1996.

[7] S. Furui, “Speaker-independent isolated word recognition using dynamic features of speech spectrum,” IEEE Trans. ASSP, vol. 34, pp.52-59, 1986.

[8] S. Furui, “On the use of hierarchial spectral dynamics in speech recognition,” Proc. ICASSP, pp. 789-792, 1990.

[9] F. Soong and M.M. Sondhi, “A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise,” IEEE Trans. ASSP, vol. 36, no. 1, pp. 41-48, 1988.

[10] J.D. Markel and A.H. Gray Jr., “Linear Prediction of Speech,” Springer Verlag, 1976.

[11] S.B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. ASSP, vol. 28, pp. 357-366, Aug. 1980.

[12] Q. Zhu and A. Alwan, “AM-Demodualtion of speech spectra and its application to noise robust speech recognition,” Proc. ICSLP, Vol. 1, pp. 341-344, 2000.

[13] A. Varga, H. Steeneken, M. Tomlinson and D. Jones, “ The NOISEX-92 study on the effect of additive noise on automatic speech recognition,” Technical report, DRA Speech Research Unit, Malvern, England, 1992.

[14] H. Dudley, “Remaking speech,” J. Acoust. Soc. Amer. 11 (2), 169-177.