On Factorizing Spectral Dynamics for Robust Speech Recognition

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

IDIAP–RR 03-33

June 2003

to appear in
Proceedings of EUROSPEECH 2003, Geneva, Switzerland

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

On Factorizing Spectral Dynamics for Robust Speech Recognition

Vivek Tyagi Iain McCowan Hervé Bourlard Hemant Misra

June 2003

to appear in
Proceedings of EUROSPEECH 2003, Geneva, Switzerland

Abstract. In this paper, we introduce new dynamic speech features based on the modulation spectrum. These features, termed Mel-cepstrum Modulation Spectrum (MCMS), map the time trajectories of the spectral dynamics into a series of slow and fast moving orthogonal components, providing a more general and discriminative range of dynamic features than traditional delta and acceleration features. The features can be seen as the outputs of an array of band-pass filters spread over the cepstral modulation frequency range of interest. In experiments, it is shown that, as well as providing a slight improvement in clean conditions, these new dynamic features yield a significant increase in speech recognition performance in various noise conditions when compared directly to the standard temporal derivative features and RASTA-PLP features.

1 Introduction

To improve the performance of automatic speech recognition (ASR) in noisy environments, increased efforts are being made towards reducing the sensitivity of ASR systems to mismatches between training data and speech data received during actual operation.

Speech is a dynamic acoustic signal with many sources of variation. As noted by Furui [4, 5], rapid 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 [1]. Therefore, representations and recognition algorithms that better use the information based on the specific temporal properties of speech should be more noise robust [2, 3]. Temporal derivative features [4, 5] of static spectral features like filter-bank, Linear Prediction (LP) [7] , or mel-frequency cepstrum [8] have yielded significant improvements in ASR performances. Similarly, the RASTA processing [2] and cepstral mean normalization (CMN) techniques, which perform cepstral high-pass 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 [1, 9]. For example, “modulation spectrum” [1] 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).

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 Section 2, we first give an overview and visualisation of the modulation frequency response. The proposed MCMS dynamic features are then derived in Section 3. Finally, Section 4 compares the performance of the MCMS features directly with standard temporal derivative features and RASTA-PLP 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

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

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

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 $X[n, \ k]$.

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

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

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

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.

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

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. Figure 1 shows an example modulation spectrum of clean speech. The figure 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 $q \in [1, \ 6]$, which corresponds to the modulation frequency range, [0Hz, 50Hz].

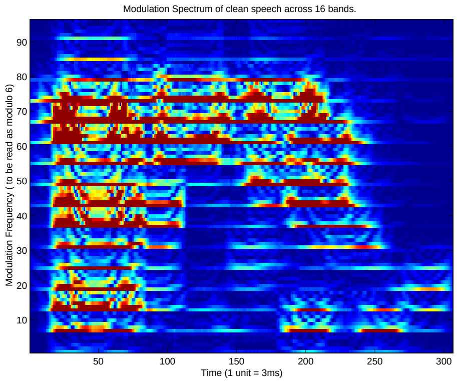

Figure 1: *Modulation Spectrum across 16 bands for a clean speech utterance. The above figure is equivalent to 16 modulation spectrums corresponding to each of 16 bands. To see q^{th} modulation frequency sample of b^{th} band, go to number $(b-1)*6+q$ on the modulation frequency axis.*

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.

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,

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

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 $M^{'}[n,\,k,\,q]$.

To compare these dynamic features with standard delta and acceleration features, Figure 2 shows trajectories of the zeroth cepstrum C_0 and its first and second temporal derivatives for a given utterance, while Figure 3 shows trajectories of the zeroth cepstrum C_0 and its third and fourth MCMS coefficients. As can be seen, the MCMS trajectories for different coefficients vary at different rates, illustrating the fact that they carry orthogonal information.

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 such as filtering operation. Figure 4 shows the cepstral modulation frequency response of the filters corresponding to first and second order derivatives of the MFCC features, while

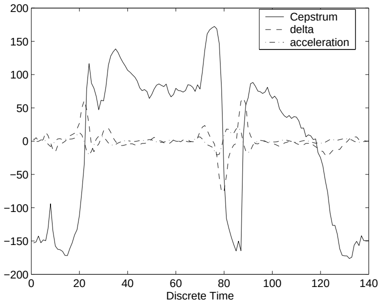

Figure 2: *Trajectories of zeroth cepstral coefficient and its first and second derivatives.*

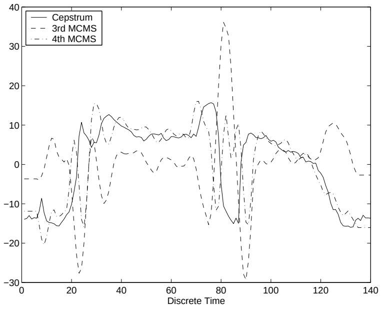

Figure 3: *Trajectories of zeroth cepstrum coefficient and its 3rd and 4th MCMS coefficient. Note that each trajectory is showing transitions at different rates. These are believed to be complementary sources of information.*

Figure 5 shows 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 and have a relatively wider band-width. This is in contrast to the MCMS features, which emphasize different cepstral modulation frequency components and have relatively narrower band-width. This further illustrates the fact that the different MCMS features carry complementary information.

4 Recognition Experiments

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

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

RASTA-PLP: 39 element feature vector consisting of 13 PLP Cepstrum and their derivatives which have been RASTA processed for noise robustness.

MFCC+MCMS: 39 element feature vector consisting of 13 MFCCs (including 0^{th} cepstral coefficient) with their 3^{rd} and 4^{th} MCMS dynamic features with variance normalization.

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. In clean conditions the baseline system gives a word error rate (WER) of 6.6%, while the MCMS system shows a slight improvement with a WER of 6.1%.

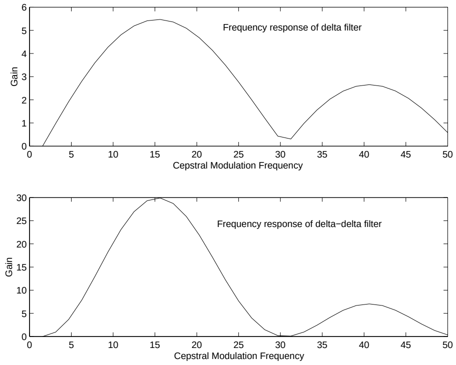

Figure 4: *Cepstral Modulation Frequency responses of the filters used in computation of derivative and acceleration of MFCC features*

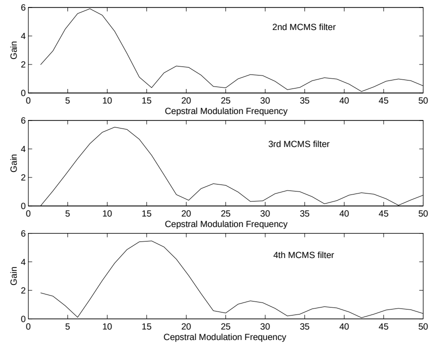

Figure 5: *Cepstral Modulation Frequency responses of the filters used in computation of MCMS features*

To verify the robustness of the features to noise, the clean test utterances were corrupted using Factory and Lynx noises from the Noisex92 database [10]. The results for the baseline and MCMS systems in various levels of noise are given in Tables 1 and 2, and plotted in Figures 6 and 7.

From these results, it is apparent that the MCMS dynamic features yield significantly greater noise robustness than standard temporal derivative features. MCMS yields comparable robustness to RASTA-PLP while providing significant improvement over RASTA-PLP in clean conditions. While in these experiments we have only used 2 MCMS coefficients (specifically, the 3^{rd} and 4^{th} coefficients) to allow a direct comparison with delta and acceleration features, in general the MCMS provides a greater range of dynamic features focused on different cepstral modulation frequencies. Further work will investigate the importance and potential of the full range of MCMS features. As these dynamic features are extracted using an orthogonal basis, the coefficients contain complementary information.

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

SNR	MFCC+Deltas	RASTA PLP	MFCC+MCMS
Clean	6.62	8.500	6.10
12 dB	23.10	14.84	15.76
6 dB	48.80	28.07	28.03

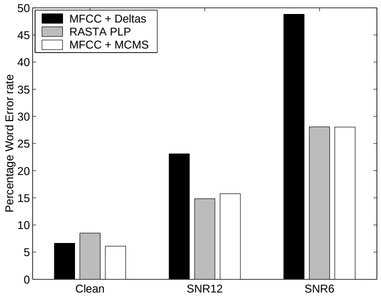

Figure 6: *Performance of MCMS features as compared to MFCC delta delta and RASTA PLP features for factory noise.*

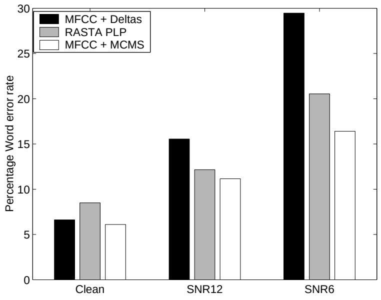

Figure 7: *Performance of MCMS features as compared to MFCC delta delta and RASTA PLP features for Lynx noise.*

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 experiments, the proposed MCMS dynamic features are compared directly to standard delta and acceleration temporal derivative features. Recognition results demonstrate that the MCMS features lead to significant performance improvement in non-stationary noise, while importantly achieving comparable performance in clean conditions. In future, we will comprehensively examine the importance of different MCMS features and will compare them with other noise robust features.

6 Acknowledgements

The authors would like to thank Prof. Hynek Hermansky of OGI, USA for his insightful comments on this work. First author would like to thank Todd Stephenson and Shajith Ikbal for the discussion with them. The authors also wish to thank DARPA for supporting this work through the EARS (Effective, Affordable, Reusable Speech-to-Text) project.

References

[1] 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.

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

SNR	MFCC+Deltas	RASTA-PLP	MFCC+MCMS
Clean	6.62	8.50	6.10
12 dB	15.55	12.16	11.16
6 dB	29.46	20.54	16.40

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

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

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

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.