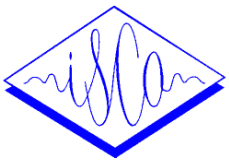

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

An Analysis of Cepstral-Time Matrices for Noise and Channel Robust Speech Recognition

Ben. P. Milner* Saeed. V. Vaseghi**

School of Information Systems, University of East Anglia, Norwich, UK.
*Now with B.T. Laboratories, Martlesham Heath, Ipswich, UK.
**Now with Queen's University Belfast.

Abstract

This paper presents an analysis of the cepstral-time matrix. The coefficients of the cepstral-time matrix are found to be similar to the standard cepstral vector with differential features augmented on. It is also shown that the cepstral-time matrix is inherently robust to convolutional channel distortion. Spectral-subtraction, Wiener filtering and model combination are extended into two-dimensions where improved noise robustness is achieved. Experimental results using the NOISEX database with noise and channel distorted speech are presented.

1. Introduction

A well-known deficiency of HMMs is the lack of an efficient mechanism for the utilisation of the correlation of successive speech feature vectors. The left-right HMM provides a temporal structure for modeling the time evolution of speech spectral characteristics from one state into the next, but within each state the observation vectors are assumed to be independent and identically distributed (IID) processes. In reality speech spectral vectors are highly correlated, and the IID assumption contributes to a rapid deterioration in the performance of feature vectors in noisy conditions. The most common method of encoding the temporal characteristics of time-varying spectral features is to extend the feature vector to include the 1st and 2nd order time derivatives. An alternative method for including the transitional information is to use a cepstral-time matrix, [1].

An analysis of the cepstral-time matrix is described in section 2, including a comparison of the columns of the cepstral-time matrix with differential cepstra. Section 3 shows how the cepstral-time matrix is robust to channel distortion, and section 4 applies some noise compensation methods to the cepstral-time matrix. Experimental results are presented in section 5. Section 6 concludes the paper.

2. Cepstral-Time Matrices

A cepstral-time matrix, $c_t(n,m)$, is obtained either by applying a 2-D DCT to a spectral-time matrix, $X_t(f,k)$, or by applying a 1-D DCT to a grouping of conventional MFCC speech vectors, [1]. M log filter bank vectors are stacked together to form a spectral-time matrix $X_t(f, k)$, where t indicates the time frame, f the frequency channel and k the time vector in the matrix. The spectral-time matrix is then transformed into a cepstral-time matrix, $c_t(n, m)$, using a 2-D DCT. Since a 2-D DCT can be divided into two 1-D DCTs, an alternative implementation of the cepstral-time matrix is to apply a 1-D DCT along the time axis of a matrix consisting of M conventional MFCC vectors.

In the cepstral-time matrix, the lower index coefficients along the axis, n, represent the spectral envelope, whereas the higher coefficients represent the pitch and excitation, as for conventional MFCCs. Along the axis, m, the lower coefficients represent the long time variation of the cepstral coefficients, and the higher coefficients the short time variation. The column m=0 represents the average or d.c. level of the spectral-time matrix. Figure 1 illustrates these regions.

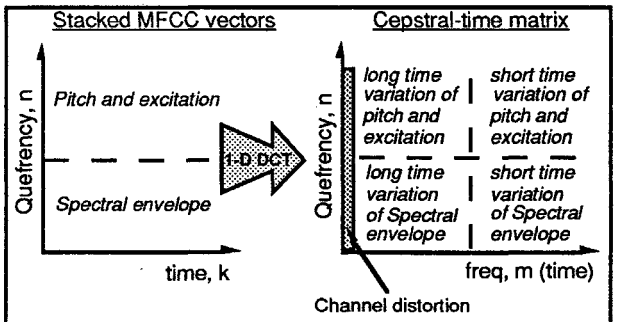

Figure 1 - Regions of cepstral-time matrix.

For speech recognition only a small part of the cepstral-time matrix is useful, thus the matrix can be truncated down to N'xM'.

2.1 Relationship between the Cepstral-Time Matrix and Differential Parameters

The cepstral-time matrix contains within it information regarding the transitional dynamics of the speech, as well as the instantaneous values. The normal way for including the speech transitional dynamics is to augment on the differential parameters [2]. A comparison can be made between the formation of the 1st and 2nd order cepstral derivatives and the columns of

the cepstral-time matrix. Equations (1) and (2) define the 1st and 2nd order cepstral derivatives

$$\partial c_t(n) = \sum_{k=-K}^{K} c_{t+k}(n)k \qquad (1)$$

$$\partial\partial c_t(n) = \partial c_{t+1}(n) - \partial c_{t-1}(n) \qquad (2)$$

It can be seen that the 1st order derivative, $\partial c_t(n)$, is produced by a weighted summation of the cepstral vectors, $c_t(n)$. The 2nd order derivative, $\partial\partial c_t(n)$ is produced by a similar summation of the 1st order derivatives, where $K=1$.

As described, the cepstral-time matrix can be obtained by applying a 1-D DCT along the time axis of a matrix containing M MFCC vectors. The second and third columns of the cepstral-time matrix, $c(n,1)$ and $c(n,2)$, are given as,

$$c(n,1) = \frac{2}{M}\sum_{k=0}^{M-1} c_k(n)\cos\frac{(2k+1)\pi}{2M} \qquad (3)$$

$$c(n,2) = \frac{2}{M}\sum_{k=0}^{M-1} c_k(n)\cos\frac{(2k+1)2\pi}{2M} \qquad (4)$$

where $c_k(n)$ is the n^{th} coefficient of the k^{th} MFCC vector in the matrix containing the stacked MFCCs. Thus it can be seen that the 1st order differential and the $c(n,1)$ column of the cepstral-time matrix are both produced by a weighted summation of the static cepstral vectors. Substituting equation (1) into (2) shows that the 2nd order differential is generated in a similar manner to that of the $c(n,2)$ column of the cepstral-time matrix. Figure 2 illustrates the similarity of basis functions for producing the differential cepstra and the second and third columns of the cepstral-time matrix.

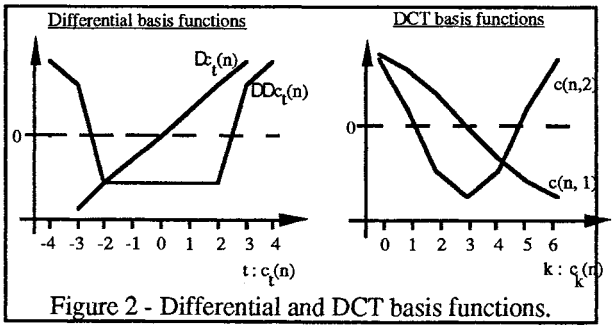

Figure 2 - Differential and DCT basis functions.

This similarity between the basis function of the DCT and differential cepstra can be extended to higher orders. Figure 3 shows the basis functions for the third and fourth differentials and the corresponding DCT basis functions.

It is interesting to note that for higher-order differential cepstra, larger time series of cepstral vectors are required. However, with the cepstral-time matrix the number of cepstral vectors required to generate higher order coefficients remains constant.

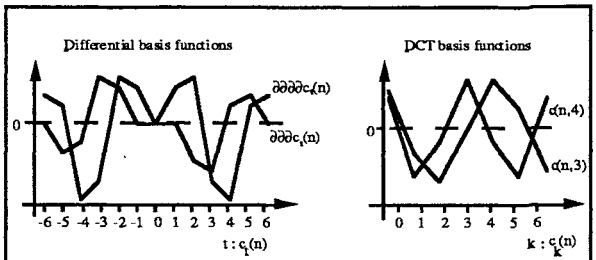

Figure 3 - 3rd and 4th order differential and DCT basis functions.

Additionally, speech transitional dynamics are produced implicitly within the cepstral-time matrix, compared to the explicit representation achieved with a cepstral vector with differential terms augmented on. Thus, cepstral-time matrices have the advantage that inverse transforms can be applied which allow transforms back into the linear filter bank domain for techniques such as parallel model combination (PMC), [5].

3. Cepstral-time matrix as a channel robust speech feature

In the frequency domain, a convolutional time-domain channel distortion is represented by a multiplicative operation,

$$Y(f) = X(f)\,H(f) \qquad (5)$$

where $Y(f)$ is the channel distorted speech, $X(f)$ is the original distortion free speech, and $H(f)$ is the frequency response of the channel. Thus, in the log filterbank domain, channel distortion becomes additive

$$\log Y(f) = \log X(f) + \log H(f) \qquad (6)$$

As the DCT is a linear transform, channel distortion is also additive in the cepstral domain,

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{h} \qquad (7)$$

where $\boldsymbol{y}_t$ is the t^{th} cepstral frame of the convolutionally channel distorted output signal and $\boldsymbol{x}_t$ is the channel input signal. The vector $\boldsymbol{h}$ represents the cepstrum of the channel, which is considered stationary, or at least slow-time varying in comparison to the channel input signal and thus has no time index. In the cepstral domain, the channel distortion, $\boldsymbol{h}$, can be considered a constant or d.c. level. Techniques such as [3] use bandpass or highpass filtering of the cepstral vectors, in time, to remove the time-invariant channel distortion, $\boldsymbol{h}$

The column m=0 of the cepstral-time matrix contains the average, or d.c. level, of the matrix of M stacked cepstral vectors. This means that the channel distortion is only present in the column, $m=0$ of the cepstral-time

the cepstral-time matrix. Equations (1) and (2) define the 1st and 2nd order cepstral derivatives

$$\partial c_t(n) = \sum_{k=-K}^{K} c_{t+k}(n)k \qquad (1)$$

$$\partial\partial c_t(n) = \partial c_{t+1}(n) - \partial c_{t-1}(n) \qquad (2)$$

It can be seen that the 1st order derivative, $\partial c_t(n)$, is produced by a weighted summation of the cepstral vectors, $c_t(n)$. The 2nd order derivative, $\partial\partial c_t(n)$ is produced by a similar summation of the 1st order derivatives, where $K=1$.

As described, the cepstral-time matrix can be obtained by applying a 1-D DCT along the time axis of a matrix containing M MFCC vectors. The second and third columns of the cepstral-time matrix, $c(n,1)$ and $c(n,2)$, are given as,

$$c(n,1) = \frac{2}{M}\sum_{k=0}^{M-1} c_k(n)\cos\frac{(2k+1)\pi}{2M} \qquad (3)$$

$$c(n,2) = \frac{2}{M}\sum_{k=0}^{M-1} c_k(n)\cos\frac{(2k+1)2\pi}{2M} \qquad (4)$$

where $c_k(n)$ is the n^{th} coefficient of the k^{th} MFCC vector in the matrix containing the stacked MFCCs. Thus it can be seen that the 1st order differential and the $c(n,1)$ column of the cepstral-time matrix are both produced by a weighted summation of the static cepstral vectors. Substituting equation (1) into (2) shows that the 2nd order differential is generated in a similar manner to that of the $c(n,2)$ column of the cepstral-time matrix. Figure 2 illustrates the similarity of basis functions for producing the differential cepstra and the second and third columns of the cepstral-time matrix.

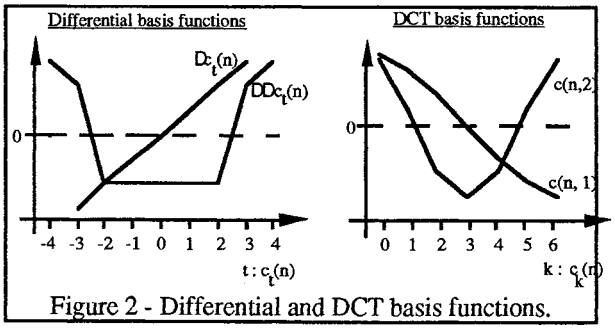

Figure 2 - Differential and DCT basis functions.

This similarity between the basis function of the DCT and differential cepstra can be extended to higher orders. Figure 3 shows the basis functions for the third and fourth differentials and the corresponding DCT basis functions.

It is interesting to note that for higher-order differential cepstra, larger time series of cepstral vectors are required. However, with the cepstral-time matrix the number of cepstral vectors required to generate higher order coefficients remains constant.

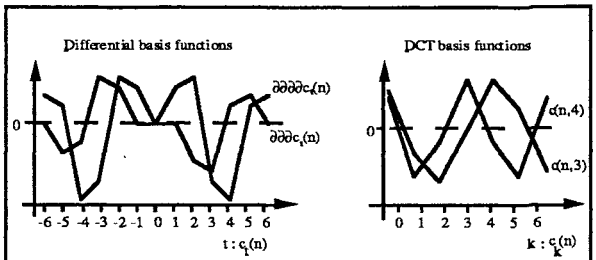

Figure 3 - 3rd and 4th order differential and DCT basis functions.

Additionally, speech transitional dynamics are produced implicitly within the cepstral-time matrix, compared to the explicit representation achieved with a cepstral vector with differential terms augmented on. Thus, cepstral-time matrices have the advantage that inverse transforms can be applied which allow transforms back into the linear filter bank domain for techniques such as parallel model combination (PMC), [5].

3. Cepstral-time matrix as a channel robust speech feature

In the frequency domain, a convolutional time-domain channel distortion is represented by a multiplicative operation,

$$Y(f) = X(f)\,H(f) \qquad (5)$$

where $Y(f)$ is the channel distorted speech, $X(f)$ is the original distortion free speech, and $H(f)$ is the frequency response of the channel. Thus, in the log filterbank domain, channel distortion becomes additive

$$\log Y(f) = \log X(f) + \log H(f) \qquad (6)$$

As the DCT is a linear transform, channel distortion is also additive in the cepstral domain,

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{h} \qquad (7)$$

where $\boldsymbol{y}_t$ is the t^{th} cepstral frame of the convolutionally channel distorted output signal and $\boldsymbol{x}_t$ is the channel input signal. The vector $\boldsymbol{h}$ represents the cepstrum of the channel, which is considered stationary, or at least slow-time varying in comparison to the channel input signal and thus has no time index. In the cepstral domain, the channel distortion, $\boldsymbol{h}$, can be considered a constant or d.c. level. Techniques such as [3] use bandpass or highpass filtering of the cepstral vectors, in time, to remove the time-invariant channel distortion, $\boldsymbol{h}$

The column m=0 of the cepstral-time matrix contains the average, or d.c. level, of the matrix of M stacked cepstral vectors. This means that the channel distortion is only present in the column, $m=0$ of the cepstral-time

spectral-time matrix. The m=0 column of the cepstral-time matrix has also been discarded.

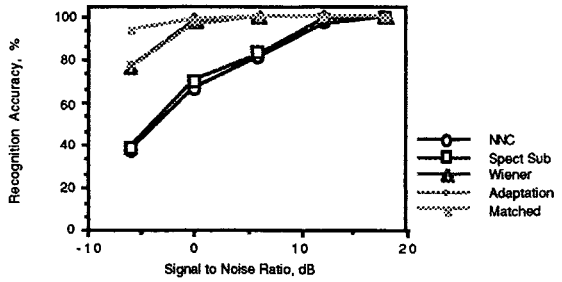

Figure 4-Recognition accuracy for helicopter noise.

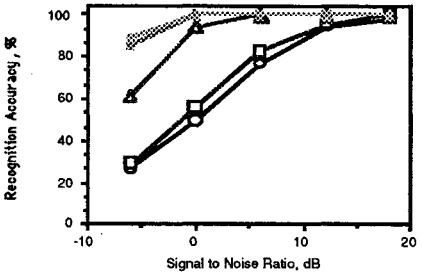

Figure 5-Recognition accuracy for car noise.

To test the channel robustness of the cepstral-time matrix, undistorted speech was used to train the HMMs. However, the channels shown in figure 6 were used to distort the test speech, thus introducing a channel-induced mismatch between the models and test speech. Table-1 shows the recognition performance for the 6 channel distortions, *a* to *f*, and for no channel distortion, *flat*. The results show the performance of a 15x4 untruncated cepstral-time matrix, and a 14x3 truncated cepstral-time matrix. Additionally the performance of 15 dimensional MFCCs is also shown. The 14x3 cepstral-time matrix remains unaffected by the channel distortions, *a* to *e*, in contrast to the 15 dimensional cepstral vector and untruncated 15x4 cepstral-time matrix which suffer degradation as a result of channel distortion. It is interesting to note that the truncated cepstral-time matrix, which is insensitive to the invertable channel distortions *a-e*, suffers some degradation as a result of the non-invertable bandpass channel response, *f*.

Matrix Dim.	*Flat*	*a*	*b*	*c*	*d*	*e*	*f*
MFCC	100	95	50	10	10	10	25
15x4 CTM	100	91	74	45	10	12	58
14x3 CTM	100	100	100	100	100	100	90

Table 1 - Recognition rate with cepstral-time matrices.

6. Conclusion

This paper has analysed the cepstral-time matrix, and shown that each column of the cepstral-time matrix contains information similar to that contained in the differential cepstrum. However the cepstral-time matrix offers an improved mathematical framework for including speech dynamics. It is also shown that the cepstral-time matrix is inherently robust to channel distortion, and that conventional noise compensation methods can be successfully extended to the cepstral-time matrix.

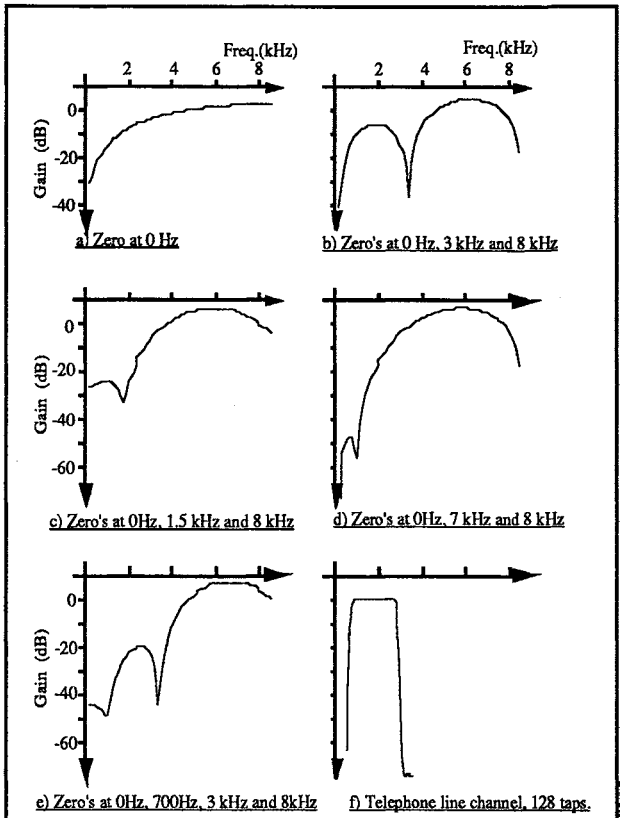

Figure 6 : Illustration of channel distortions

7. Acknowledgments

The authors wish to acknowledge the support of BT Laboratories and the Science and Engineering Research Council.

8. References

[1] B.P. Milner and S.V. Vaseghi, "Comparison of some noise-compensation methods for speech recognition in adverse environments", Proc. IEE, vol. I, pp601-604, Oct. 1994.

[2] B.A. Hanson and T.H. Applebaum,"Robust speaker-independent word features using static, dynamic and acceleration features; experiments with Lombard and noisy speech",Proc. ICASSP-90, pp857-860.

[3] H. Hermansky *et al*, "RASTA-PLP speech analysis technique", Proc. ICASSP-92, pp121-124.

[4] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. ASSP-27, pp113-120, 1979.

[5] M.J.F. Gales and S.J. Young, "HMM recognition in noise using parallel model combination", Proc. EuroSpeech-93, pp837-840.