Mel Frequency Cepstral Coefficients for Music Modeling

Beth Logan
Cambridge Research Laboratory
Compaq Computer Corporation
One Cambridge Center
Cambridge MA 02142
Beth.Logan@compaq.com

Abstract

We examine in some detail Mel Frequency Cepstral Coefficients (MFCCs) - the dominant features used for speech recognition - and investigate their applicability to modeling music. In particular, we examine two of the main assumptions of the process of forming MFCCs: the use of the Mel frequency scale to model the spectra; and the use of the Discrete Cosine Transform (DCT) to decorrelate the Mel-spectral vectors.

We examine the first assumption in the context of speech/music discrimination. Our results show that the use of the Mel scale for modeling music is at least not harmful for this problem, although further experimentation is needed to verify that this is the optimal scale in the general case. We investigate the second assumption by examining the basis vectors of the theoretically optimal transform to decorrelate music and speech spectral vectors. Our results demonstrate that the use of the DCT to decorrelate vectors is appropriate for both speech and music spectra.

Keywords: Music representation, music features, MFCC features.

1 Introduction

Of all the man-made sounds which influence our lives, speech and music are arguably the most prolific. Speech has long been perceived as a natural interface between people and computers and hence has received much focused attention. Decades of research in the speech community has led to usable systems and convergence of the features and models used for speech analysis.

In the music community however, although the field of synthesis is very mature, a dominant paradigm has yet to emerge to solve other problems such as music classification or transcription. Consequently, many representations for music have been proposed (e.g. (Martin, Scheirer & Vercoe 1998), (Martin 1998), (Scheirer & Slaney 1997), (Blum, Keislar, Wheaton & Wold 1999)). In this paper, we examine some of the assumptions of Mel Frequency Cepstral Coefficients (MFCCs) - the dominant features used for speech recognition - and examine whether these assumptions are valid for modeling music.

MFCCs are short-term spectral-based features. Therefore, for the purposes of this paper, we will assume that we are only interested in short-term spectral-based features of music. While it is clear that other features might be of use for a particular problem (e.g. rhythm or beats or any other of the features referenced above), it is also clear that the spectral composition of a signal contains much information. This information could certainly be augmented by additional features if required or accumulated over longer time windows.

MFCCs have been used by other authors to model music and audio sounds. For example, in (Foote 1997) a retrieval system is built based on a cepstral representation of sounds. Blum et. al. also list MFCCs as one of the features in their retrieval system (Blum et al. 1999). A music summarization system based on cepstral features is described in (Logan & Chu 2000). These works however use cepstral features merely because they have been so successful for speech recognition without examining the assumptions made in great detail.

The organization of this paper is as follows. We first describe MFCCs for modeling speech. We then investigate whether several of the assumptions made when forming these features are appropriate for music. Finally we present conclusions and suggestions for future work.

2 MFCCs for Speech Recognition

MFCCs have been the dominant features used for speech recognition for some time (e.g. (Young, Woodland & Byrne 1993)). Their success has been due to their ability to represent the speech amplitude spectrum in a compact form. Each step in the process of creating MFCC features is motivated by perceptual or computational considerations. We examine these steps in more detail in the following paragraphs. A more complete description of the

process and assumptions is given in (Rabiner & Juang 1993).

Figure 1 shows the process of creating MFCC features. The first step is to divide the speech signal into frames, usually by applying a windowing function at fixed intervals. The aim here is to model small (typically 20ms) sections of the signal that are statistically stationary. The window function, typically a Hamming window, removes edge effects. We generate a cepstral feature vector for each frame.

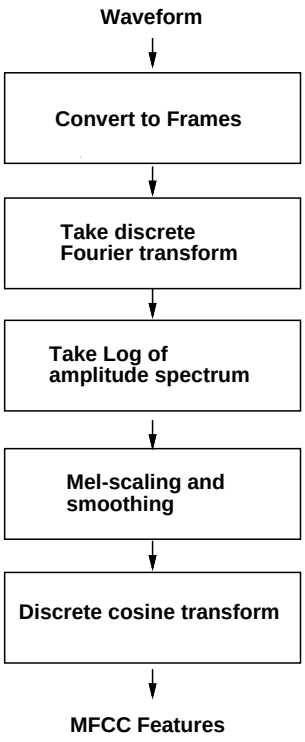

Figure 1: Process to create MFCC features

The next step is to take the Discrete Fourier Transform (DFT) of each frame. We then retain only the logarithm of the amplitude spectrum. We discard phase information because perceptual studies have shown that the amplitude of the spectrum is much more important than the phase. We take the logarithm of the amplitude spectrum because the perceived loudness of a signal has been found to be approximately logarithmic.

The next step is to smooth the spectrum and emphasize perceptually meaningful fre-

quencies. This is achieved by collecting the (say) 256 spectral components into (say) 40 frequency bins as shown in Figure 2. Although one would expect these bins to be equally spaced in frequency, it has been found that for speech, the lower frequencies are perceptually more important than the higher frequencies. Therefore, the bin spacing follows the so-called 'Mel' frequency scale.

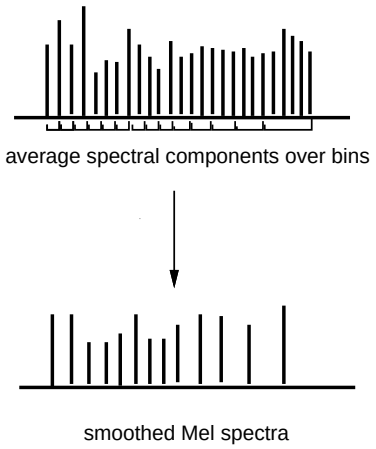

Figure 2: Mel scaling and smoothing of the log amplitude spectrum. Spectral components are averaged over Mel-spaced bins to produce a smoothed spectrum.

The Mel scale is based on a mapping between actual frequency and perceived pitch as apparently the human auditory system does not perceive pitch in a linear manner. The mapping is approximately linear below 1kHz and logarithmic above. Figure 3 shows the Mel function.

The components of the Mel-spectral vectors calculated for each frame are highly correlated. Speech features are typically modeled by mixtures of Gaussian densities. Therefore, in order to reduce the number of parameters in the system, the last step of MFCC feature construction it to apply a transform to the Mel-spectral vectors which decorrelates their components. Theoretically, the Karhunen-Loeve (KL) transform (or equivalently Principal Components Analysis (PCA)) achieves this. In the speech community, the KL transform is approximated by the Discrete Cosine Transform (DCT) (Marhav & Lee 1993). Using this transform, 13 (or so) cepstral features are obtained for each frame.

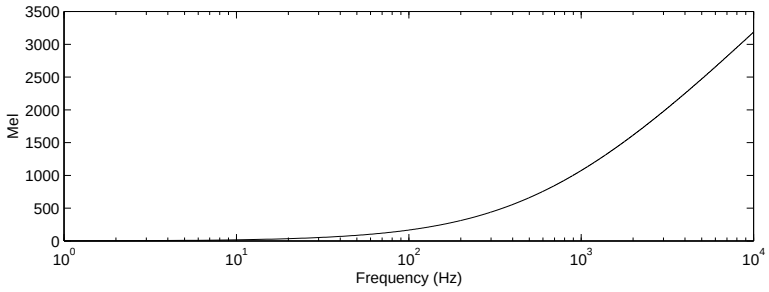

Figure 3: The Mel scale.

3 MFCCs for Music Analysis

As discussed above, the process of calculating MFCCs for speech consists of five main steps.

1. Divide signal into frames.
2. Obtain the amplitude spectrum.
3. Take the logarithm.
4. Convert to Mel spectrum.
5. Take the DCT.

We now seek to determine whether this process is suitable for creating features to model music. Of the steps listed above, we shall examine steps 4 and 5. The other steps are less controversial since music, like speech is non-stationary (step 1) and phase-independent (for mono[1] recordings) (step 2). Also, although music has a larger dynamic range than speech, there is no reason to suspect that loudness will not be perceived logarithmically (step 3).

However, it is possible that the Mel spectra may not be optimal for music signals as there may be more information in say higher frequencies (step 4). Similarly, the DCT may not approximate the KL transform for music (step 5). We shall therefore examine these assumptions in the following sections.

[1] as opposed to stereo

3.1 Mel vs Linear Spectral Modeling

To investigate whether it is appropriate to determine the spectrum of music using the Mel scale, we examine the performance of a simple speech/music discriminator. We have available around 3 hours of labeled in-house data from a broadcast news show. The show contains interviews and commercials and has a number of segments of music. The data is divided into 2 hours of training data and 40 minutes of testing data. Around 10% of the training data and 14% of the testing data is labeled as music.

We convert the training data to 'Mel' and 'Linear' cepstral features as follows. The signal is sampled at 16kHz and converted to 25.6ms frames overlapped by 10ms. For the 'Mel' features, we then convert each frame to 40 Mel-spaced frequency components. For the 'Linear' features, a linear frequency scale is used. Referring to Figure 2, this corresponds to using Mel-spaced bins in the first case and linearly-spaced bins in the second. In both cases, we then take the logarithm and DCT as usual to obtain 13 cepstral features for each frame.

Using a standard version of the Expectation-Maximization (E-M) algorithm (e.g. (Baum, Petrie, Soules & Weiss 1970)), we train mixture of Gaussian classifiers for the labeled speech and music segments in the training data. Our Gaussian densities have diagonal covariance matrices since we assume that our 13 cepstral components are uncorrelated. We then classify each segment in the test data by the model - speech or music - which has the highest average likelihood over the test segment.

Table 1 shows the segmentation error for models with varying number of mixture components built with the two different features types. It shows that for this speech/music classification problem, the results are (statistically) significantly better if Mel-based cepstral features rather than linear-based cepstral features are used. However, whether this is simply because the Mel scale models speech better or because it also models music better is not clear. It is also not clear whether a different frequency warping and/or a larger sampling rate would be beneficial. At worst, we can conclude that using the Mel cepstrum to model music in this speech/music discrimination problem is not harmful. Further tests are needed to verify that the Mel cepstrum is appropriate for modeling music in the general case.

3.2 Using the DCT to Approximate the KL Transform

We now investigate the effectiveness of using the DCT to decorrelate Mel spectral features. As mentioned in Section 2, the correct way to decorrelate components is to use the KL transform. We will therefore first investigate how well the DCT approximates the KL transform for speech spectra.

The KL transform converts vector $\mathbf{u}$ of dimension m to vector $\mathbf{v}$ of dimension n where $n \leq m$ and the components of $\mathbf{v}$ are uncorrelated (e.g. (Ghanem 1991)). This can be

Number Mix.	Spectrum	Segmental Error (%)
4	Mel	7.1
	Linear	14.3
8	Mel	1.8
	Linear	8.9*
16	Mel	3.6
	Linear	10.7*

Table 1: Classification results for testing data tested with models with varying number of mixture components and different types of spectra. * indicates result is significantly worse than the result immediately above with 95% confidence using a Matched-Pairs test.

expressed as

$$\mathbf{v} = \mathbf{O}\mathbf{u}. \tag{1}$$

Here $\mathbf{O}$ is the nxm transformation matrix. It can be shown that the components of $\mathbf{v}$ will be uncorrelated if the rows $\mathbf{O}$ are the orthonormalized eigenvectors of the covariance matrix $\mathbf{R}$ of $\mathbf{u}$.

The eigenvalues of $\mathbf{R}$ rate the importance of each corresponding eigenvector. For speech signals, typically $n < m$ eigenvalues are significant implying that each feature vector $\mathbf{u}$ can be transformed to vector $\mathbf{v}$ of smaller dimension (e.g. (Ephraim & VanTrees 1995)). This redundancy exists because the rank of $\mathbf{R}$ is less than m, implying that dependencies exist between the components of $\mathbf{u}$. Thus typically, we convert a Mel-spectral vector of dimension 40 to a cepstral vector of dimension 13.

Figure 4 shows the eigenvalues and first 15 eigenvectors of $\mathbf{R}$ for a sequence of Mel log spectral vectors collected from about 3 hours of speech. The speech used is the training set for TIMIT (Garofolo et al. 1993) which is a speaker independent, clean speech database. The decay of the eigenvalues is clearly seen implying that only the first 10 or so are significant.

The DCT which is used in the speech community to approximate the KL transform can be written as

$$\mathbf{c} = \mathbf{D}\mathbf{u}. \tag{2}$$

Here the elements of $\mathbf{c}$ are the cepstral coefficients of $\mathbf{u}$ and $\mathbf{D}$ is a nxm matrix of cosine basis functions (i.e. cosine functions).

Figure 5 shows the first 15 cosine basis functions. Comparing Figure 4 with this figure we see that the eigenvector-derived basis functions of the KL transform are 'cos-like' in nature, particularly for the more important first few functions with larger eigenvalues. This

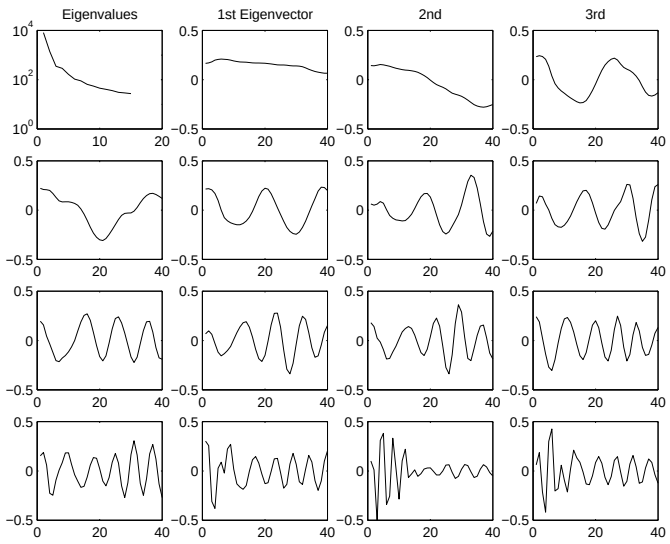

Figure 4: Eigenvalues and first 15 eigenvectors for the covariance matrix of Mel log spectral vectors for 3 hours of clean speech

observation is also noted in (Hermansky & Malayath 1998).[2] Thus we have demonstrated that the DCT is close to the optimal decorrelation function for speech log spectra.

We now examine the KL transform for music spectra. Figure 6 shows the eigenvalues and first 15 eigenvectors of $\mathbf{R}$ for a sequence of Mel log spectral vectors collected from 100 approximately 3 minute Beatles songs (289 minutes of music total). Again, we see that the eigenvalues decay rapidly and that the basis functions are 'cos-like' in nature, again particularly the first few. Thus we conclude that the use of the DCT for decorrelating music log spectra is appropriate.

[2]The order of KL basis functions shown is slightly different to their corresponding cosine functions. The cosine functions are shown in order of increasing frequency argument. The KL functions are ordered by their corresponding eigenvalue which does not necessarily give exactly the same order. In particular note that the 3rd and 4th eigenvectors of the speech data appear to be 'reversed' compared to the cosine basis functions. Actually, the eigenvalues for these two eigenvectors are almost the same number as shown on the eigenvalue plot.

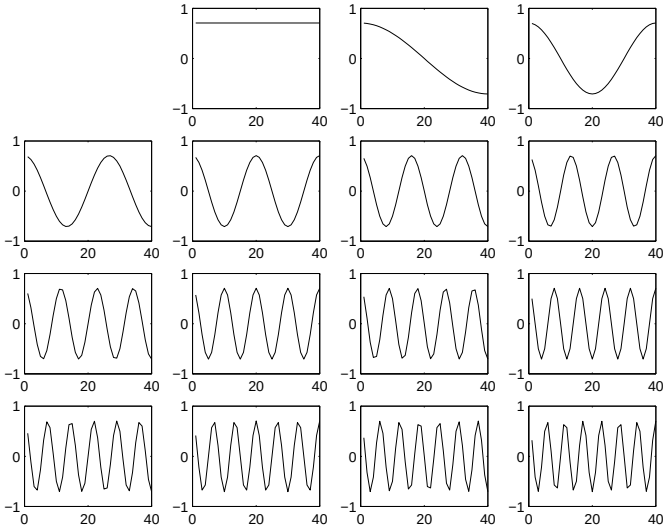

Figure 5: First 15 cosine basis functions

4 Conclusions and Suggestions for Future Work

In this paper, we have sought to build on the success in the speech recognition community by investigating how applicable it is to use the dominant features for modeling speech to model music. We first discussed the process of forming MFCC features for speech, describing the reasons for and assumptions made at each step. We then investigated two of the more controversial steps in the context of music modeling.

First, we examined the use of the perceptual Mel frequency scale in the context of speech/music discrimination. We found that the Mel scale was at least not harmful for this problem, although further experimentation is needed to verify that this is the optimal scale for modeling music spectra in the general case.

Second, we investigated the assumption that the last step of forming cepstral features results in decorrelated vectors. By examining the basis vectors of the theoretically optimal transform to decorrelate vectors, we demonstrated that this assumption is appropriate both for speech and music spectra.

Future work should focus on a more thorough examination of the spectral parameters to use such as the sampling rate of the signal, the frequency scaling (Mel or otherwise) and the number of bins to use when smoothing. Also worthy of investigation is the windowing size and frame rate.

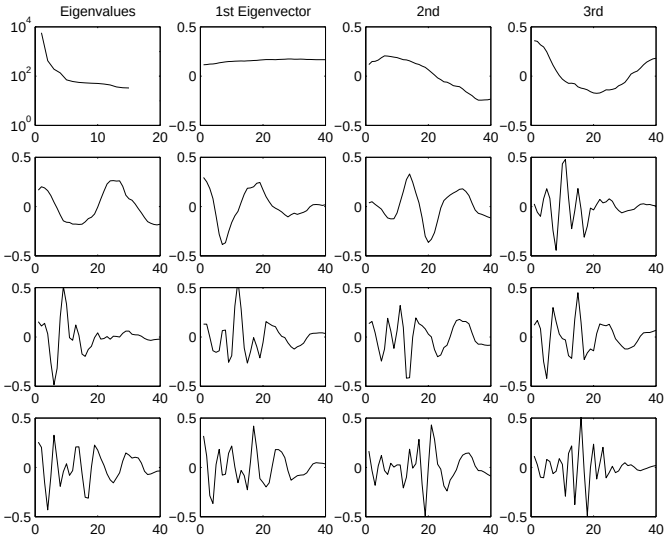

Figure 6: Eigenvalues and first 15 eigenvectors for the covariance matrix for Mel log spectral vectors for about 4 1/2 hours of Beatles songs

5 Acknowledgments

The author would like to thank the speech group at CRL for helpful discussions and moral support.

References

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970), 'A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains', *Ann. Math. Statist.* **41**, 164–171.

Blum, T. L., Keislar, D. F., Wheaton, J. A. & Wold, E. H. (1999), *Method and article of manufacture for content-based analysis, storage, retrieval, and segmentation of audio information*, U.S. Patent 5, 918, 223.

Ephraim, Y. & VanTrees, H. L. (1995), 'A signal subspace approach for speech enhancement', *IEEE Transactions on Speech and Audio Processing* **3**(4), 251–265.

Foote, J. T. (1997), Content-based retrieval of music and audio, *in* 'SPIE', pp. 138–147.

Garofolo, J. S. et al. (1993), Darpa timit. acoustic-phonetic continuous speech corpus. nistir 4930, Technical report, DARPA.

Ghanem, R. G. (1991), *Stochastic finite elements: A Spectral Approach*, Springer-Verlag.

Hermansky, H. & Malayath, N. (1998), 'Spectral basis functions from discriminant analysis', *International Conference on Spoken Language Processing.*

Logan, B. T. & Chu, S. (2000), Music summarization using key phrases, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing'.

Marhav, N. & Lee, C.-H. (1993), 'On the asymptotic statistical behavior of empirical cepstral coefficients', *IEEE Transactions on Signal Processing* **41**, 1990–1993.

Martin, K. D. (1998), Toward automatic sound source recognition: identifying musical instruments, *in* 'Proc. NATO Computational Hearing Advanced Study Institute'.

Martin, K. D., Scheirer, E. D. & Vercoe, B. L. (1998), Music content analysis through models of audition, *in* 'Proc. ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications'.

Rabiner, L. R. & Juang, B. H. (1993), *Fundamentals of Speech Recognition*, Prentice–Hall.

Scheirer, E. & Slaney, M. (1997), Construction and evaluation of a robust multifeature speech/music discriminator, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing'.

Young, S. J., Woodland, P. C. & Byrne, W. J. (1993), *HTK: Hidden Markov Model Toolkit V1.5*, Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc.