Mel Frequency Cepstral Coefficients for Music Modeling

Beth Logan
Cambridge Research Laboratory
Compaq Computer Corporation
One Cambridge Center
Cambridge MA 02142
USA
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.

MFCCs for Music Analysis

Of all the human generated sounds which influence our lives, speech and music are arguably the most prolific. Speech has received much focused attention and decades of research in this community have led to usable systems and convergence of the features 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. (Martin1998), (Scheirer1997), (Blum1999)). 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 have been used by other authors to model music and audio sounds (e.g. (Blum1999)). These works however use cepstral features merely because they have been so successful for speech recognition without examining the assumptions made in great detail.

MFCCs (e.g. see (Rabiner1993)) are short-term spectral features. They are calculated as follows (the steps and assumptions made are explained in more detail in the full paper):

1. Divide signal into frames.
2. For each frame, obtain the amplitude spectrum.
3. Take the logarithm.
4. Convert to Mel (a perceptually-based) spectrum.
5. Take the discrete cosine transform (DCT).

We seek to determine whether this process is suitable for creating features to model music. We examine only steps 4 and 5 since, as explained in the full paper, the other steps are less controversial.

Step 4 calculates the log amplitude spectrum on the so-called Mel scale. This transformation emphasizes lower frequencies which are perceptually more meaningful for speech. It is possible however that the Mel scale may not be optimal for music as there may be more information in say higher frequencies. Step 5 takes the DCT of the Mel spectra. For speech, this approximates principal components analysis (PCA) which decorrelates the components of the feature vectors. We investigate whether this transform is valid for music spectra.

Mel vs Linear Spectral Modeling

To investigate the effect of using the Mel scale, we examine the performance of a simple speech/music discriminator. We use around 3 hours of labeled data from a broadcast news show, divided into 2 hours of training data and 40 minutes of testing data. We convert the data to ‘Mel’ and ‘Linear’ cepstral features and train mixture of Gaussian classifiers for each class. We then classify each segment in the test data using these models. This process is described in more detail in the full paper.

We find 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. 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.

Using the DCT to Approximate Principal Components Analysis

We additionally investigate the effectiveness of using the DCT to decorrelate Mel spectral features. The mathematically correct way to decorrelate components is to use PCA (or equivalently the KL transform). This transform uses the eigenvalues of the covariance matrix of the data to be modeled as basis vectors. By investigating how closely these vectors approximate cosine functions we can get a feel for how well the DCT approximates PCA. By inspecting the eigenvectors for the Mel log spectra for around 3 hours of speech and 4 hours of music we see that the DCT is an appropriate transform for decorrelating music (and speech) log spectra.

Future Work

Future work should focus on a more thorough examination the parameters used to generate MFCC features 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.

Suggested Readings

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

Martin, K.. 1998, Toward automatic sound source recognition: identifying musical instruments, Proceedings NATO Computational Hearing Advanced Study Institute.

Rabiner, L. and Juang, B., 1993, Fundamentals of Speech Recognition, Prentice-Hall.

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