The Use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification

Róisín Loughran
University of Limerick, Limerick, Ireland

Jacqueline Walker
University of Limerick, Limerick, Ireland

Michael O'Neill
University College Dublin, Dublin, Ireland

Marion O'Farrell
University of Limerick, Limerick, Ireland

ABSTRACT

This paper examines the use of Mel-frequency Cepstral Coefficients in the classification of musical instruments. 2004 piano, violin and flute samples are analysed to get their coefficients. These coefficients are reduced using principal component analysis and used to train a multi-layered perceptron. The network is trained on the first 3, 4 and 5 principal components calculated from the envelope of the changes in the coefficients. This trained network is then used to classify novel input samples. By training and testing the network on a different number of coefficients, the optimum number of coefficients to include for identifying a musical instrument is determined. We conclude that using 4 principal components from the first 15 coefficients gives the most accurate classification results.

1. INTRODUCTION

The human ability to distinguish between musical instruments has been a subject of investigation for a number of years. Even with minimal musical exposure, most people can easily distinguish between familiar musical instruments, even when played at the same loudness and pitch. By definition [1] that quality of auditory sensation by which a listener can distinguish between two sounds of equal loudness, duration and pitch is known as timbre. Hence it could be said that musical instrument recognition is largely dependent on timbre. Unfortunately, unlike pitch and loudness, timbre has proven to be somewhat difficult to measure or quantify.

In the past, speech analysis has dominated the field of audio research and consequently received more attention than its musical counterpart. It is not surprising then, that many researchers in musical analysis would look to the features and methods employed in speech analysis when examining musical tones. This paper examines one such feature. Mel-frequency Cepstral Coefficients (MFCC) have been used extensively in speech analysis over the past few decades [2] and have more recently received attention in music analysis [3]. This paper tries to distinguish between musical instruments using only MFCCs and looks at how many of these coefficients are necessary and useful for accurate instrument identification. Section 2 discusses some previous work in the area of musical sound identification and in particular, studies involving MFCCs. Section 3 outlines the proposal and methods used in this study. Section 4 describes the results obtained and finally Section 5 outlines our conclusion and proposes further work in this area.

2. PREVIOUS WORK

Research to determine and distinguish between different classes of instruments has become more popular as the field of audio analysis has expanded more into music analysis. Herrera et al [4] give quite an exhaustive review of methods used in automatic identification of musical instruments. From this review it is evident that both temporal and spectral qualities are needed for accurate instrument identification. The current study looks at the use of MFCCs - a spectral quality, over the temporal duration of the note.

A number of studies have looked to MFCCs in sound identification. De Poli and Prandoni [5] used MFCCs in their study of timbre space. Brown [6] distinguished between oboes and saxophone sounds by calculating cepstral coefficients and applying a k-means algorithm to form clusters. Eronen and Klapuri [7] included MFCCs as one of their features in examining a wide range of orchestral instruments. Logan [8] examines some of the finer points of the MFCC in music analysis as opposed to speech analysis and determined that it is indeed useful in this domain.

3. PROPOSAL

When using MFCCs in speech analysis, it has been determined that 8-14 coefficients are sufficient to use and quite often 12 are chosen [2]. Although MFCCs have been used in music identification, there has been no such recommendation for this purpose. Hence in this study the aim is to determine how many coefficients are suitable for musical sound identification. This is implemented using Principal Component Analysis (PCA) to reduce the dimensionality of the data and then this reduced data is used to train a Multi-Layered Perceptron (MLP).

3.1. Mel-frequency Cepstral Coefficients

MFCCs are a way of representing the spectral information in a sound. Each coefficient has a value for each frame of the sound. The changes within each

coefficient across the range of the sound are examined here. Obtaining the MFCCs involves analysing and processing the sound according to the following steps [8]:

1. Divide the signal into frames
2. Get the amplitude spectrum of each frame
3. Take the log of these spectrums
4. Convert to the Mel scale
5. Apply the Discrete Cosine Transform (DCT)

The Mel scale is a perceptual scale that is based on human hearing. The DCT in step 5 actually approximates the PCA (described later) in that it reduces the data orthnormally, thus leaving a series of uncorrelated values (the coefficients) for each frame of the sound. Hence after this algorithm has been used, we are left with a matrix of values for each sample sound that is the number of coefficients by the number of frames in size. This is implemented in Matlab using the melcepst function from the voicebox toolbox [9].

These calculated coefficients change from frame to frame. The changes in these values can be plotted as an envelope across the sound. These envelopes are distinctive to the instrument as illustrated below. *Figure 1* shows the changes in the first MFCC for C5 on a piano whereas *figure 2* shows the first MFCC for the same note on the flute. These changes are examined here as a method of identifying the instruments.

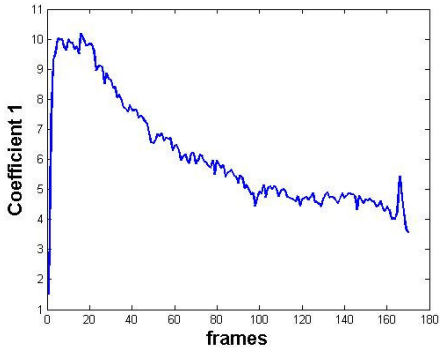

Figure 1 Trend of the first MFFC for C5 on a piano

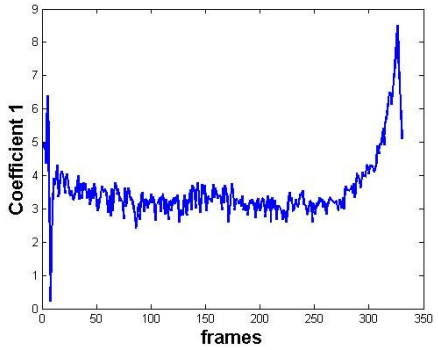

Figure 2 Trend of the first MFCC for C5 on a flute

3.2. Principal Component Analysis

The measures discussed above will all have multiple data points per envelope. Statistically, much of this data is redundant and so a method to extract the most significant information from the data collected must be determined. This is achieved through applying PCA to the calculated coefficient data. PCA is a standard technique commonly used in statistical pattern recognition and signal processing for performing dimensional reduction. This was implemented in Matlab for the experiment using the princomp function in the Statistics Toolbox. Essentially it transforms data orthonormally so that the variance of the data remains constant, but is concentrated in the lower dimensions. The matrix of data being transformed consists of one set of coefficients for each sample. Thus there is now one matrix of data for each cepstral coefficient. The covariance matrix of the data matrix is then calculated. The principal components for the data set can be calculated from the eigenvectors of this covariance matrix [10]. This results in a set of principal components, with variance ordered from highest to lowest. As such the most important data is extracted with minimum disruption to the original data collected. While this method may not leave particularly intuitive or meaningful data axes, it is an excellent method of reducing the calculated data. Graphically, up to three principal components are easy to plot and visualise, although less significant components may still contain significant data.

3.3. Multi-layered Perceptron

MLPs are a specific type of Artificial Neural Network (ANN) that use supervised training to train multiple layers of interconnected perceptrons. MLPs contain at least one layer of hidden neurons – each of which includes a non-linear activation function, and they exhibit a high degree of connectivity [11]. These characteristics combine to make the theoretical analysis of an MLP difficult and as such the design of these systems is often, as in this case, unintuitive and based on trial and error. The network used in this experiment is trained using the backpropagation algorithm with two hidden layers of neurons.

3.4. Data Sets

3.4.1. Training Data

It was decided for this study to exhaustively search just three instruments – the piano, violin and flute. Samples were taken from the RWC Music Database (Music Instrument Sound) of these 3 instruments. Three makes of piano, Yamaha, Bosendorfer and Steinway were each sampled at dynamic levels *f*, *mf* and *p* across their range [12]. Violins manufactured by J.F Pressenda, Carcassi and Fiumebianca were sampled at these three loudness levels with vibrato and at level *mf* without vibrato across their range [13].

Plucked violin samples were not incorporated into this dataset. Flutes manufactured by Louis Lot and Sankyo were sampled at the three levels both with and without vibrato [14]. In total this gave 2004 samples across the entire pitch range of the three instruments.

3.4.2. Test Data

The samples that make up the test dataset are from the MUMS (McGill University Master Samples) database [15]. This smaller database consists of samples of the three instruments played at the same dynamic level. In total this dataset consists of 45 violin samples, 37 flute samples and 88 piano samples. Each instrument was sampled and recorded across their entire range. A completely different dataset from the training set was used, as this should test the generality of the classifier.

4. RESULTS

The results from this experiment rely on looking at the changes in the MFCCs across the length of the notes of each instrument and using these changes as a way of recognising the instrument. Once the principal components of each coefficient were calculated, the first three components can be plotted to observe the separation between the instruments. One such plot for the second coefficient can be seen in *figure 3*. Here clustering of each instrument can be observed. Similar such plots can be created for the other MFCCs.

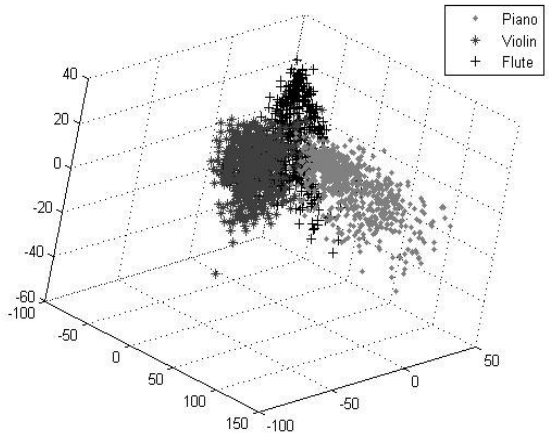

Figure 3 Plot of the first 3 principal components of MFCC2 for the 3 instruments

Once the data had been reduced and the principal values extracted, these values were used to train a MLP. The MLP was implemented in Matlab using the *newff* function from the Neural Network Toolbox. This was set up with a learning rate of 0.1 and a momentum constant of 0.95. It is batch trained, with a goal of 0.001 and trained up to maximum epochs of 400. With this set up it was found that a network with 50 neurons in the first layer and two hidden layers containing 18 and 15 neurons respectively would be sufficient to train the data set. The MUMS test data is then used to simulate the network and the results are given as the percentage of times the trained network recognises these sounds correctly.

4.1. Using the first 3 Principal Components

Initially the first 3 principal components calculated were examined. Preliminary results indicated that unless at least six MFCCs were used the results were not encouraging. Hence the network was trained with the first 6 to 16 MFCCs to compare the classification results. Each training and testing set was run 10 times. The average of these test results can be seen in *Figure 4*.

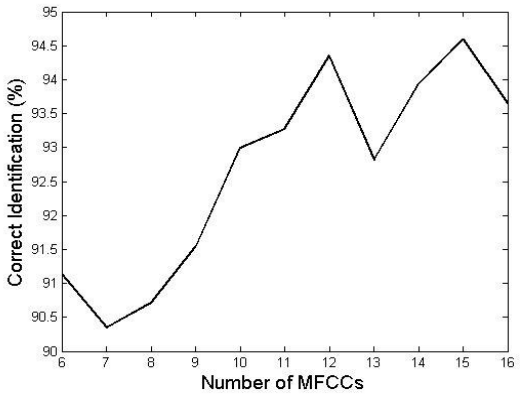

Figure 4 Classification results for network trained on first 3 principal components

These results indicate that once more than 10 MFCCs are used, the recognition results are consistently high. Using 15 MFCCs does give the highest recognition rate, but the more values incorporated the more computationally expensive the calculations become.

4.2. Using more Principal Components

It is also worth considering that using more principal components for each coefficient may lead to even more accurate results. Although it is not possible to plot more than three dimensions, the MLP can accept more than three input principal components for each coefficient. Test results for a network trained on the first three, four and five principal components of the first 11 to 16 coefficients are shown in the bar chart below in *figure 5.*

These results clearly indicate that, in general, using 4 principal components increases the accuracy of the classification. Including the 5th actually reduces the result. This may be due to the unintuitive way in which the PCA reduces data. It is unclear what physical aspect, if any, each component depends on and so it is possible that this 5th one is dependent on a frequency or dynamic element of the sound and not on the instrument. From this bar chart it can be seen that using the first 4 principal components from 15 MFCCs gives a classification result of 95.88%. This is quite a high and

encouraging result as this classifier is based only on MFCCs.

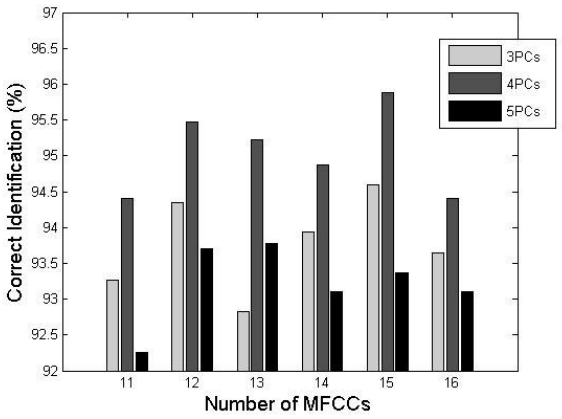

Figure 5 Comparison of results for different number of principle components

5. CONCLUSION

This paper examined the use of MFCCs in musical instrument recognition. Examining this with PCA and MLPs, it was possible to discern the optimum number of MFCCs to include for classification and how many principal components of these coefficients to apply. From the results we can conclude that at least 10 MFCCs should be used. It was observed that taking 4 principal components gave the best classification and that the highest result was obtained from using 15 MFCCs.

This classifier only looks at one specific measure of a sound – the MFCCs, and yet still achieves quite accurate results. To improve the standard of this classifier even further more spectral and temporal features of the sounds need to be included. Now that we have decided on our optimum data set from MFCCs we can combine this with these other features to create a more robust classifier. We would also look at other classifier methods. As mentioned, the MLP offer somewhat of a 'black box' solution to out problem and so we may look to other types of Neural Networks such as an ARTMAP [16] to give us more control over the system.

6. ACKNOWLEDGEMENT

This study is funded by the Science Federation Ireland (SFI) under the current National Development Plan and Strategy for Science Technology and Innovation (SSTI) 2006-2013.

7. REFERENCES

[1] ASA, Acoustical Terminology, New York: American Standards Association, New York (1960)

[2] O'Shaughnessy, D.: Speech Communication Human and Machine, Addison-Wesley Series in Electrical Engineering (1987)

[3] Eronen, A.: Musical Instrument Recognition Using ICA-Based Transform of Features and Discriminatively Trained HMMS. In ISSPA (2003)

[4] Herrera, Pamatriain, X., Batlle, E., Serra, X.: Towards Instrument Segmentation for Music Content Description: A Critical View of Instrument Classification Techniques. In ISMIR (2000)

[5] De Poli, G., Prandoni, P.: Sonological Models for Timbre Characterization. J. New Music Research, **26**, 170-197 (1997)

[6] Brown, J.: Computer Identification of Musical Instruments Using Pattern Recognition with Cepstral Coefficients as Features. J. Acoust. Soc. Am. **105**, 1933-1941 (1998)

[7] Eronen, A., Klapuri, A.: Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 753-756 (2000)

[8] Logan, B.,: Mel Frequency Cepstral Coefficients for Music Modelling. In ISMIR, (2000)

[9] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[10] http://www1.cs.columbia.edu/~jebara/htmlpapers/UTHESIS/node64.html

[11] Haykin, S.: Neural Networks A Comprehensive Foundation. Prentice Hall International (UK) Limited, London (1999)

[12] RWC Music Database: RWC-MDB-I-2001-W01, Instrument No.1: Pianoforte

[13] RWC Music Database: RWC-MDB-I-2001-W05, Instrument No.15: Violin

[14] RWC Music Database: RWC-MDB-I-2001-W09, Instrument No.33: Flute

[15] http://www.music.mcgill.ca/resources/mums/html/mums.html

[16] Carpenter, G.A., Grossberg, S., Reynolds, J.H.: ARTMAP: Supervised Real-Time Learning and Classification of Nonstationary Data by a Self-Organizing Neural Network. In: Neural Networks, **4**, 565-588 (1991)