Music Instrument Identification Using MFCC:

Erhu as an Example

Chih-Wen Weng, Cheng-Yuan Lin**, Jyh-Shing Roger Jang***

*Chinese Music Dept, Tainan National College of The Arts, Taiwan
Computer Science Dept, Tsing Hua University, Taiwan
Email: ivanweng@giga.net.tw
**Computer Science Dept, Tsing Hua University, Taiwan
Email: {gavins, jang}@wayne.cs.nthu.edu.tw

Abstract：

In the analysis of musical acoustics, we usually use the power spectrum to describe the difference between timbres from two music instruments. However, according to our experiments, the power spectrum cannot be used as effective features for erhu instrument identification. In this paper, we use MFCC (mel-scale frequency cepstral coefficients) as features for music instrument identification using GMM (Gaussian mixture models); the result is very encouraging. MFCC and GMM are commonly used in speech/speaker recognition with success. This paper demonstrates that MFCC and GMM can also be used for erhu instrument identification. Immediate extension of the current work includes MFCC-based music instrument assessment and music acoustic analysis.

Keywords: Timbre, Power Spectrum, MFCC, GMM, Music Instrument Identification, Erhu Music Instrument

1. Introduction

Conventional analysis of music timbre is usually based on power spectrum to find the frequency distribution of a given audio music signal. In this paper, we propose the use of MFCC (mel frequency cepstral coefficients) [12][15] for timbre classification and music instrument identification. This study is motivated by the success of speech and speaker recognition using MFCC, which is a non-parametric method modeling the human auditory perception system. Our experiments involve the use of GMM (Gaussian mixture models) for music instrument identification of erhu

(an instrument for traditional Chinese music), using both the features of power spectrum and MFCC. The initial results demonstrate the feasibility of using MFCC for such purpose.

The rest of the paper is organized as follows. Section 2 explains the design method for our ETR (erhu timber recognition) system. Section 3 covers the experimental results and corresponding discussion. Section 4 gives conclusions and possible future directions of this study.

2. ETR System Design

In the following, we shall describe how to build an erhu timbre recognition (ETR for short) system. Similar to most pattern recognition systems (such as speaker recognition [1][7]), ETR involves the following steps:

1. Feature extraction: How to extract discriminant features from the given recordings.
2. Model construction: How to construct the ETR system from the extracted features in the previous step.
3. Model application: How to use the constructed model for classifying new recordings.

For most speech and speaker recognition systems, there are several methods for feature extraction, including power spectrum, formant, LPC (linear predictive coding), LSP (line spectrum pair) and MFCC (mel frequency cepstral coefficients) [12][15]. Power spectrum is mostly used to explain the difference in timbres of two instruments. On the other hand, MFCC is by far the most commonly used feature for speech and speaker recognition. Therefore in this paper, we shall investigate the use of power spectrum and MFCC for ETR and compare their performance.

Several frequently used classifiers for model construction in pattern recognition include GMM (Gaussian mixture models) [4], VQ (vector quantization) [16], ANN (artificial neural networks) [11], SVM (support vector machines) [14], linear classifiers [13], and so on. Our past experiments for speech and speaker recognition indicate that GMM can usually achieve a better recognition rate. Therefore, we use GMM as the classifier for our ETR system. The following figure indicates the flowchart of our ETR system.

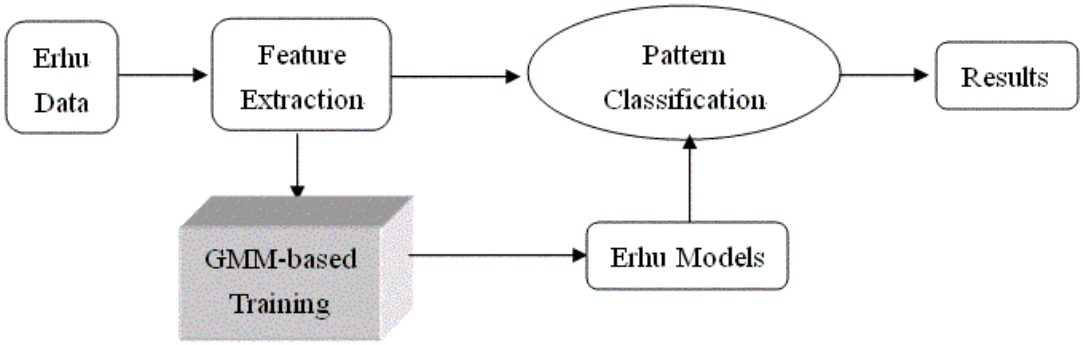


Figure 1. The ETR system's flowchart.

Power spectrum is frequently used for explaining the timbre for music instruments. In our experiment, we use fast Fourier transform (FFT) to extract the power spectrum of each frame from the recording of erhu. Each frame consists of 512 sample points, leading to a feature dimension of 256 for power spectrum. This dimension is too large for most model construction methods. Therefore, we choose LDA (linear discriminant analysis) [8] algorithm to select the most discriminant dimensions. The LDA algorithm is a statistical technique for data classification and dimensionality reduction in pattern recognition. The principal concept of LDA is to identify a new basis that can maximize the ratio of between-class variance to the within-class variance. In this paper, we keep 24 discriminant dimensions of each power spectrum for next model construction/application.

The MFCC (mel frequency cepstral coefficients) [9] is a very popular feature in speech and speaker recognition. It can be derived using the following steps:

1. Pre-emphasis
2. Hanning (or Hamming) windowing
3. FFT to obtain power spectrum
4. Triangular bandpass filtering
5. Discrete cosine transform to obtain MFCC
6. Taking delta MFCC (optional)

The use of MFCC has been justified by the satisfactory performance of speaker/speech recognition systems available in the literature. Consequently, we choose MFCC for our ETR system.

GMM (Gaussian mixture model) is a classic parametric model used in many

pattern recognition applications. The GMM assumes that the probability distribution of the observed data takes the following form:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{m} \alpha_i N(\vec{x}; \mu_i, \Sigma_i), \quad (1)$$

where $N(\vec{x}; \mu_i, \Sigma_i)$ denotes the p-dimensional normal distribution with mean vector μ and covariance matrix Σ, defined by

$$N(\vec{x}; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2}} |\Sigma_i|^{-1/2} \exp\left[-\frac{1}{2} (\vec{x} - \mu_i)^T \Sigma_i^{-1} (\vec{x} - \mu_i) \right] \quad . \quad (2)$$

In (1) the terms α_i are positive scalar weights ($\sum_{i=1}^{m} \alpha_i = 1 \quad and \quad \alpha_i \geq 0$) and $\vec{x}$ is a p-dimensional vector. In this paper, $\vec{x}$ is actually a p-dimensional vector representing the features of each frame. Each erhu instrument is represented by a GMM and is referred to by its model λ_i.

$$\lambda_i = \{\alpha_i, \mu_i, \Sigma_i\}, \quad i = 1,....,10 .$$

GMM can have several different forms depending on the choice of covariance matrices. The covariance matrix has three different types, including one covariance matrix per Gaussian component (nodal covariance), one covariance matrix for all Gaussian components in an erhu model (grand covariance) or a single covariance matrix shared by all erhu models. In addition, the covariance matrix can also be full or diagonal. In this paper, nodal and diagonal covariance matrices are primarily used for erhu modeling.

There are several methods for estimating the parameters of GMM. So far the maximum likelihood (ML) [6] estimation has been the most well-established one. The goal of ML estimation is to derive the optimum model parameters that can maximize the likelihood of GMM. Generally, the model parameters are extracted by Expectation-Maximization (EM) [2] algorithm. This algorithm is also used in estimating the HMM (hidden Markov models) parameters, which is also known as the Baum-Welch reestimation algorithm [10].

The statistical models of T erhu instruments can be denoted by $\{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_T\}$. In this paper, the value of T is 10. We apply the following equation to find the erhu model which has the maximum a posteriori probability for a given observation sequence X:

$$N = \arg\max_{1 \leq k \leq T} \Pr(\lambda_k \mid X) = \arg\max_{1 \leq k \leq T} \frac{p(X \mid \lambda_k) \Pr(\lambda_k)}{p(X)}$$

Assuming equally likely erhu instruments ($\Pr(\lambda_k)$ =1/T) and noting that $p(X)$ is the same for all erhu models, the classification rule simplifies to

$$N = \arg\max_{1 \le k \le T} p(X \mid \lambda_k)$$

And using the logarithms and the independence between observations, we have the following equation:

$$N = \arg\max_{1 \le k \le T} \sum_{j=1}^{J} \log p(\vec{x}_j \mid \lambda_k),$$

Where $\vec{x}_j$ is the feature vector for j-th frame.

3. Experimental Results

For the following experiment, we have collected 11 recordings from each of 10 different erhu instruments. These 110 files were recorded when one of the authors played these 10 erhu instruments with different music, speed, volume, style, and so on. For each erhu, we use its 10 recordings as the training data and the remainder one as the test data. The recording conditions are listed next:

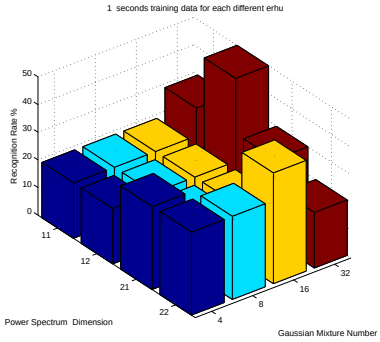
- Duration: 12 seconds
- Sampling rate: 16 KHz
- Bit resolution: 16 bits
- Number of channels: one (mono)

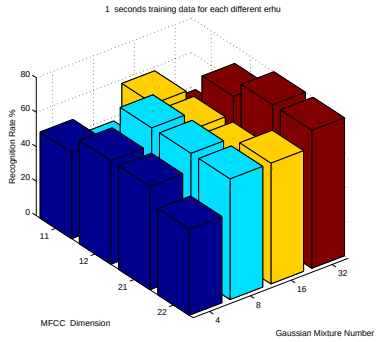
For efficiency consideration, we only try three different durations: 1, 3, 5 seconds for both training and test. For experimental parameters, we have combined 4 different numbers of feature dimensions and GMM mixture counts, as shown next:

- The feature dimensions of power spectrum or MFCC: 11, 12, 22 and 24. For power spectrum, the selected features are obtained from LDA. For MFCC, the selected features are:
  - 11: Original MFCC, excluding the first term
  - 12: Original MFCC
  - 22: Derivative of the 11 features mentioned above
  - 24: Derivative of the 12 features mentioned above
- The mixture counts of GMM: 4, 8, 16 and 32.

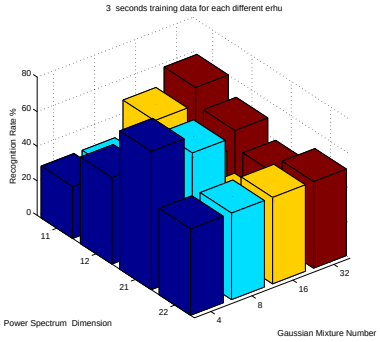
The following two diagrams show the recognition rates using power spectrum and MFCC, respectively, when the duration is only 1 seconds. It is obvious that MFCC can achieve better recognition rates than power spectrum for all the 16 cases. The average recognitions are 26.25% and 61.875% for power spectrum and MFCC, respectively. (The recognition rates referred in the following discussion are based on

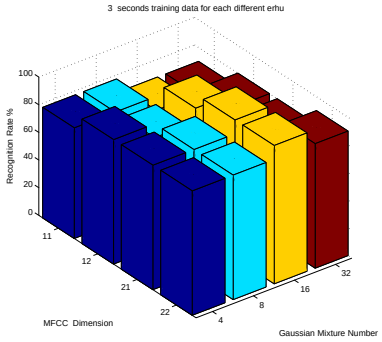
the leave-one-out method.)



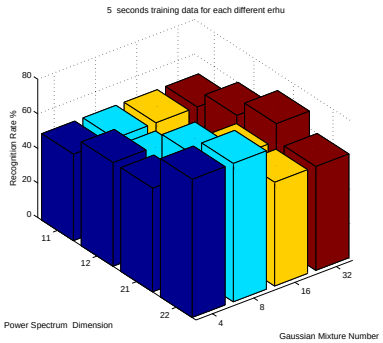


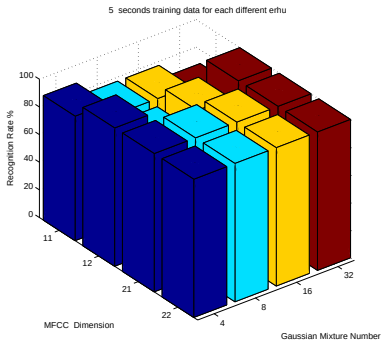
The following two diagrams show the recognition rates when the duration is 3 seconds. Again, MFCC can achieve better recognition rates than power spectrum for all the 16 cases. The average recognitions are 48.75% and 85.625% for power spectrum and MFCC, respectively.





The following two diagrams show the recognition rates when the duration is 5 seconds. Again, MFCC can achieve better recognition rates for all the 16 cases. The average recognitions are 59.375% and 93.125% for power spectrum and MFCC, respectively. In particular, when the feature dimension is 22 or 24, ETR based on MFCC can achieve 100% recognition rates for all GMM mixture counts.





From the above experimental results, it is obvious that MFCC is more effective than power spectrum. Moreover, as the duration is 5 seconds or longer, the recognition rates are close to 100%. This indicates MFCC is a feasible feature for

erhu instrument identification.

4. Conclusions and Future Work

This paper provides an empirical study of using MFCC and power spectrum for erhu music instrument identification. The experiments demonstrate that MFCC is a much more effective feature for this purpose. This is justifiable since MFCC is widely used in speech/speaker recognition and it is aimed to model the human auditory perception system.

An immediate future work of this study is to use other audio features, such as PLP [15] or Wavelet Transform [3]. Moreover, different classes of instruments have different timbres. Therefore we can apply the same algorithm for music instrument assessment, which can provide important information for music performers and instrument manufacturers.

5. Reference

[1] "Automatic recognition of speakers from their voices," Proc. IEEE, vol.64 pp. 460-475, 1976

[2] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Stat. Soc., vol. 39, pp. 1-38, 1977.

[3] C.S. Burrus, R.A. Gopinath, and h. Guo, Introduction to wavelets and Wavelet Transform, Prentice-hall International, Inc., New Jersey, 1998

[4] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Processing, vol. 3, no. 1, pp. 72-83, 1995.

[5] Duhamel, P. and M. Vetterli, "Fast Fourier Transforms: A Tutorial Review and a State of the Art," Signal Processing, Vol. 19, April 1990, pp. 259-299.

[6] G. McLachuo, Mixture Models, New York: Marcel Dokker, 1998.

[7] G. R. Doddington, "Speaker recognition – identifying people by their voices," Proc. IEEE, vol. 73, pp. 1651-1644, Nov. 1985.

[8] J. Duchene and S. Leclercq, "An Optimal Transformation for Discriminant Principal Component Analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 10, No 6, November 1988

[9] Krishnamurthy, A.K. and D.G. Childers, "Two Channel Speech Analysis," IEEE Trans. on Acoustics, Speech and Signal Processing, 1986, 34, pp. 730-743.

[10] L. Baum et al., "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann, Math Stat., vol. 41, pp. 164-171, 1970.

[11] L. Fausett, *Fundamentals of Neural Networks,* New Jersey, Prentice-Hall 1994.

[12] Rabiner, Juang, "Fundamentals of speech recognition", published by Prentice Hall.

[13] Richard O. Duda, Peter E. Hart, David G. Stork *Pattern classification* (2nd edition), Wiley, New York, 2001

[14] Saunders, C., Stitson, M. O., Weston, J., Bottou, L., Schoelkopf., B. and Smola, A. (1998) Support Vector Machine - Reference Manual. Technical Report CSD-TR-98-03, Department of Computer Science, Royal Holloway, University of London.
[15] Xuedong Huang, Alex Acero, H. W. Hon, “Spoken language processing”, published by Prentice Hall.
[16] Y. Linde, A. Buzo, & R. M. Gray, “An Algorithm for Vector Quantization Design,” *IEEE Transactions on Communication*, v. COM-208, pp.84-95, 1980.

Authors’ Biography

Chih-Wen Weng is a full-time lecturer in Chinese Music department at National Tainan Art University, Taiwan, since 1996. He won the first prize of the erhu category in 1987’s music competition in Taiwan. He has also received several music composition prizes. His research interests include music temperament analysis and computer music synthesis. Since 2003, he has been a PhD candidate in the Department of Computer Science at National Tsing Hua University, Taiwan.

Cheng-Yuan Lin received the Master degree in Department of Computer Science at National Tsing Hua University, Taiwan, in 2001. Since 2003, he has been a PhD candidate in the same department. His research interests include speech/singing/music synthesis, speaker/speech/music recognition and audio signal processing.

Jyh-Shing Roger Jang received the Ph.D. degree in EECS Department at UC Berkeley in 1992. Since 1995, he has been with the Department of Computer Science, National Tsing Hua University, Taiwan. His research interests include melody/music/speech/speaker recognition, audio signal processing/recognition, pattern recognition, neural networks and fuzzy logic.