MUSIC INSTRUMENT RECOGNITION : FROM ISOLATED NOTES TO SOLO PHRASES

Krishna A.G and T.V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore-560012, India
email: *krishna@protocol.ece.iisc.ernet.in* and *tvsree@ece.iisc.ernet.in*

ABSTRACT

Speech and Audio processing techniques are used along with statistical pattern recognition principles to solve the problem of music instrument recognition. Non temporal, frame level features only are used so that the proposed system is scalable from the isolated notes to the solo instrumental phrases scenario without the need for temporal segmentation of solo music. Based on their effectiveness in speech, Line Spectral Frequencies(LSF) are proposed as features for music instrument recognition. The proposed system has also been evaluated using MFCC and LPCC features. Gaussian Mixture Models and K-Nearest Neighbour model classifier are used for classification. The experimental dataset included the UIowa's MIS and the C Music corporation's RWC databases. Our best results at the instrument family level is about 95% and at the instrument level is about 90% when classifying 14 instruments.

1. INTRODUCTION

While much has been achieved in the field of speech content analysis(Automatic Speech Recognition (ASR), Language Identification (LID), Speaker Identification (SID) etc), music content analysis is relatively in its infancy. In the broad area of Music Content Analysis, sound source recognition (recognition of musical instruments and others) forms a very important part. Music Content analysis has a lot of applications including media annotation, singer identification, music transcription, structured audio coding, information retrieval etc. Drawing analogies from speech processing, ASR corresponds to automatic music transcription, LID to music genre recognition and SID to music instrument recognition. The solution to these three important problems has reached a certain maturity in speech, and we look to draw from that, although speech and music are quite different.

There has been a lot of work in the area of Music Instrument Recognition (MIR). A brief collection of those that are most relevent to the work presented here are discussed. Brown [2] has used SID techniques to determine the properties most useful in identifying sounds from 4 woodwind instruments. Cepstral coefficients, bin-to-bin differences of constant-Q transform coefficients and autocorrelation coefficients were used as features with gaussian mixture model based classifiers, obtaining accuracies of about 79% to 84%. Excerpts from commercial CDs were used in her study rather than isolated notes. Marques [3] has used gaussian mixture models and support vector machines to classify 0.2s segments of 9 instruments obtaining an accuracy of about 70% with LPC, FFT based cepstral coefficients and MFCC feature sets. Marques also has used solo music, and not isolated notes.Martin [1] has used a set of perceptual features derived from a lag-log correlogram to classify isolated notes from 27 instruments with accuracies of about 86% at the instrument family level and about 71% at the individual instrument level. This system has been shown to be robust with respect to handling noisy and reverberent notes. Eronen [4] has also used a set of perceptually motivated features to classify isolated notes from 30 instruments with accuracies of about 94% at the instrument family level and about 85% at the individual instrument level. Agostini [6] has used spectral features only to classify 27 instruments with an accuracy of about 96% at the instrument family level and about 92% at the individual instrument level. Eronen [5], in a study of comparing different features for music instrument recognition has reported best accuracies of 77% at the instrument family level and 35% at the individual instrument level. The different feature sets analyzed are LPCC, on an uniform as well as warped frequency scale, MFCC and other features. The best accuracies were obtained for WLPCC with a prediction order of 13 and Bark scale warping. Kitahara [7] has classified tones from 19 instruments using a fundamental frequency dependent multivariate normal distribution of spectral, temporal, modulation and other features obtaining accuracies of about 90% at the instrument family level and about 80% at the individual instrument level. Except Brown and Marques, all other results are using isolated notes.

In this paper, we propose the use of Line Spectral Frequencies, an alternative representation of the conventional

0-7803-8484-9/04/$20.00 ©2004 IEEE

LPC, for music instrument recognition. Based on the success of speaker identification, one can use frame level features to classify musical instruments. However, in the case of speech, it is easy to get clean speech for speaker enrolment; but, in the case of music, real life music has, often, multiple instruments. Hence, keeping to the same approach, we develop models on isolated notes (equivalent to clean speech) and we are extending the experiment to solo music phrases.

2. FEATURE EXTRACTION

The current study is aimed at the recognition of musical instruments from either isolated notes or solo music phrases. Energy thresholding is done to extract the quasi steady state portion of the notes and frame level features derived from these portions alone are used for recognition. A parameter α controls the degree of thresholding. A zero value for α is equivalent to extracting the whole note as such without discarding any portion of it. While it is beyond doubt that several explicit temporal features like onset time etc play a big role in music instrument recognition, extracting such features in real world music is a difficult task. In real world music(or even solo music) defining exact points of attack, decay etc is an ill-posed problem. In practice we would prefer a system that works on both isolated notes and solo music irrespective of whether it was trained on isolated notes or solo music. To achieve this robustness we have avoided explicit temporal features such as rise time etc. Thus, our approach is towards scalability. If explicit temporal properties are to be used on solo or real world music, then a preprocessing stage which does very fine, reliable and consistent temporal segmentation is needed.

The issue of scalability requires robust features. We propose the use of Line Spectral Frequencies (LSFs) as robust features for music instrument recognition. While LSFs are used quite successfully in speech coding, recognition and enhancement, they have not been used for music instrument recognition. Yet, LSFs are known to be more robust and amenable for perceptual weighting of the feature components.

In LPC analysis, a short segment of the signal is assumed to be generated from an all pole filter $H(z) = 1/A(z)$, where $A(z)$ is given by

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_M z^{-M} \tag{1}$$

Here M is the order of the LPC analysis and the filter coefficients are the LPC coefficients. To define the LSFs, the inverse filter polynomial is used to construct the following two augmented polynomials,

$$P(z) = A(z) + z^{-(M+1)} A(z^{-1}) \tag{2}$$

$$Q(z) = A(z) - z^{-(M+1)} A(z^{-1}) \tag{3}$$

The roots of the polynomials $P(z)$ and $Q(z)$ are referred to as the LSFs[8]. LSFs characterize the resonances and their bandwidths of $A(z)$. There are some unique properties of LSFs, such as interleaving, which is useful for quantization and perceptual weighting. Also, spectral sensitivities of LSFs are localized ie., a change in a particular LSF produces a change in the power spectrum only in its neighborhood. Music instruments are known to have characterestic resonances (their locations as well as bandwidths) which are important determinants of timbre. The LSFs, since they model the resonances or peaks directly, are more suited for music instrument recognition than the LPCs which are sensitive to overall spectral shape and hence change quite drastically with small changes in spectral shape with no significant changes in the resonances.

To benchmark the performance of the LSF features, we used the MFCC and LPCC feature sets in our experiments. The rationale behind using MFCC for music instrument recognition is as follows. In computing the MFCC feature set, generally, the zeroth and the higher order coefficients are not considered. This can be viewed as doing away with intensity information and pitch information, which is what is needed since we want to recognize music instruments independent of the intensity and the pitch thus focussing on timbre. This is in addition to the mel-based filter spacing and dynamic range compression in the log filter outputs, which represent the human auditory system in a simplified way.

3. STATISTICAL MODELLING

Since the frame level feature vectors are obtained from different musical notes, they can be viewed as statistically independent. We have used Gaussian Mixture Models to represent each class and then performed a maximum-likelihood (ML) classification. The GMM classifier is a parametric classifier wherein the data is assumed to have a probability density function that can be modelled as a weighted sum of multivariate gaussian probability density functions. Each gaussian density is called a mixture component. A GMM with M mixtures is given as

$$p(X|\lambda) = \Sigma_{j=1}^{j=M} \omega_j \mathcal{N}(x, \mu_j, \Sigma_j) \tag{4}$$

Refer [2] for details.

Given N models, optimum classification is done by the ML rule :

$$\lambda^{\star} = argmax_{\lambda} log\{p(x|\lambda)\} \tag{5}$$

We have used the GMM on frame level features and average likelihood of all the frames belonging to a note is used to classify the note. In the case of solo music, average likelihood of all the frames of a particular duration is used.

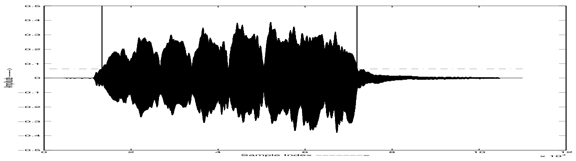

Fig. 1. Thresholding with α=1.2 on a violin note

Another non-parametric optimum classifier is the K-Nearest Neighbor classifier (KNNC) which can approach ideal performance asymptotically with training data. The classifier can provide arbitrarily complex boundaries between classes in the feature space (non-convex, disjoint etc), and is a very simple and intutively appealing classifier.

The KNNC has the drawback of having to store all the training patterns for testing which can be a burden both from a computation and storage point of view, especially when dealing with a large number of classes and large data. For a more compact representation of the frame-level feature space, we have used a different feature space for the K-NNC. In this, each note is represented by one model comprising of the mean and the diagonal covariances of the feature vectors within a note. Several notes belonging to an instrument become a set of such models. During classification, the test data frames are converted into the model space, and the K-NN in the model space determines the final decision.

4. EXPERIMENTS AND RESULTS

We first experimented on finding an optimal prediction order keeping the threshold used for extracting the quasi stationary portion from the note constant. Then, with this prediction order we obtained the optimal thresholding measure. We then experimented on finding the optimal number of mixtures for the GMM classifier and optimal K for the KNN classifier to obtain the best classification performance. The accuracies presented are all obtained by a 6-fold cross validation. We have used a rectangular window of 23ms, without pre-emphasis, and a prediction order of 20 in our experiments with a α of 0.01, giving a threshold which is 1% of the mean of the full-wave rectified signal. Figure 1 shows the thresholding with $\alpha = 1.2$ for illustration purposes. The portion inbetween the 2 vertical lines is considered for processing.

We have used the MIS database[1] for comparing the proposed LSF feature set with MFCC and LPCC features. The details of the data for our experiment are as in Table 1.

Table-2 shows the performance of the GMM classifier on the three feature sets, for different mixtures varying from 26 to 54. This range was determined experimentally. The optimal number of mixtures is data dependent, for our study, 46-50 can be considered optimal. LSFs are clearly seen to outperform the MFCC and LPCC features. The best performance with optimal training data was 95% and 90% at the instrument family and instrument level respectively. Resonances in the spectrum, their locations and Qs are more effectively and explicitly modelled by LSFs, as compared to LPCC and MFCC. In instrument recognition accuracies LPCC does better than the MFCC in the individual instrument case and in the instrument family case. This is because LPCC contains more fine structure information of the power spectrum, compared to MFCC where the power spectrum will have been smoothed.

Table-3 is similar to the previous table, except that the classifier is K-NNC. While there is not much variation in the performance of the LSF features, that of MFCC, especially in the instrument case is much improved, and that of LPCC is very slightly deteriorated. This is because the averaging done in arriving at the K-NNC feature space from the frame-level features has resulted in loss of the fine structure information which was helping the LPCC do better in the GMM case. The MFCC, on the other hand seems to have gained from the smoothing that has occured in the K-NNC feature space.

Table-4 shows the individual instrument accuracies, using GMM classifier. Violin and Cello are the best classified instruments, while the French horn is poorly classified by LSF and LPCC. In general, it was noticed that most confusions occuring were within the instrument families.

We also tested our system on the RWC database[2] with the same set of instruments as in [7]. Our best results with this database are summarized in Table-5. This performance is very good considering the simplicity of the feature set and the classifier. In addition, our feature set is easily extendible to the solo instrumental music case, since it does not rely on any explicit temporal features of isolated notes which cannot be extracted from solo instrumental music without a very fine temporal segmentation algorithm.

To demonstrate the scalability of our approach, we used short segments (about 1.2s in length each) of music from the RWC Jazz music database and performed classification using the models built from isolated notes. On a forced 3 way classification procedure, flute, piano and guitar instrumental pieces were classified with 74% accuracy by GMMs built from isolated notes, thus demonstrating the scalability of our system.

[1] University of Iowa's Music Instrument Samples,http://theremin.music.uiowa.edu/MIS.html

[2] RWC Music databases by C-Music Corporation, http://staff.aist.go.jp/m.goto/RWC-MDB/

5. CONCLUSIONS

A new approach to music instrument recognition is proposed, wherein the models built from isolated notes are shown to be useful for instrument identification from solo music phrases, without the need for temporal segmentation. The novel feature set of LSF for music instrument recognition, is shown to be superior to that of MFCC and LPCC. Future work involves developing and identifying more robust features for the task and better statistical modelling. Perceptual distance measures is also a very fertile area for research, to improve the performance of music instrument recognition systems.

Table 1. Details of the database

MIS Database
Strings : Violin, Cello
Flutes : Alto flute, Bass flute, Flute
Reeds : Bassoon, Oboe, BbClarinet, EbClarinet
Brass : French horn, Alto Sax, Tenor Trombone, Bass Trombone, Soprano Sax

Table 2. Performance of different feature sets across mixtures (GMM Classifier)(All accuracies are in percentages. Instrument accuracy is given outside and family accuracy inside paranthesis)

Mixtures	**LSF**	**LPCC**	**MFCC**
26	86.00 (92.66)	82.07 (87.00)	74.85 (83.20)
30	86.57 (92.83)	81.71 (86.96)	76.90 (86.81)
34	87.02 (93.40)	79.47 (85.06)	75.45 (85.79)
38	86.96 (93.06)	80.41 (87.30)	76.47 (84.64)
42	87.02 (93.06)	82.03 (87.80)	76.92 (86.13)
46	87.41 (93.28)	82.65 (86.70)	78.52 (86.19)
50	87.53 (93.17)	82.87 (87.64)	76.86 (84.73)
54	87.25 (93.06)	81.83 (86.59)	75.71 (83.70)

Table 3. Performance of different feature sets (K-NNC)

K	**LSF**	**LPCC**	**MFCC**
1	87.90 (91.45)	81.65 (86.81)	79.36 (86.09)
3	86.23 (91.23)	77.98 (83.47)	76.20 (85.57)
5	85.08 (90.59)	75.68 (82.77)	74.99 (85.63)

Table 4. Individual Music Instrument Recognition using LSF features and 46 mixture GMM

Instrument	**LSF**	**LPCC**	**MFCC**
Alto Sax	97.10	88.13	96.00
Bass Trombone	72.28	66.42	95.40
BbClarinet	73.43	77.62	26.50
EbClarinet	62.81	55.78	44.80
Sop Sax	85.68	83.64	79.90
Tenor Trombone	92.41	87.16	87.60
Alto Flute	85.66	87.64	56.53
Bass Flute	100	94.07	82.70
Bassoon	91.04	90.10	75.00
Cello	99.00	100.00	95.75
Flute	85.86	80.67	83.15
French Horn	55.71	62.50	75.30
Oboe	98.60	93.33	66.00
Violin	96.44	86.00	91.75

Table 5. Performance of LSF feature set on the RWC database (GMM Classifier)

Task	**Accuracy in %**
19 instr notes classification	77(84)
3 instr phrases classification	74
3 instr phrase classification with 19 targets	41

6. REFERENCES

[1] K. D. Martin, *Sound Source Recognition: A Theory and Computational Model*, Phd Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999.

[2] J. C. Brown, O. Houix and S. McAdams, *Feature dependence in the automatic identification of musical woodwind instruments*, J. Acoust. Soc. Am., Vol. 109, No. 3, pp. 1064-1072, 2001.

[3] J. Marques and P. Moreno, *A study of musical instrument classification using gaussian mixture models and support vector machines*, Cambridge Research Labs Technical Report Series CRL/4, 1999

[4] A. Eronen and A. Klapuri, *Music instrument recognition using cepstral coefficients and temporal features*, Proc. of ICASSP, pp. 753-756, 2000.

[5] A. Eronen, *Comparison of features for musical instrument recognition*, Workshop on Signal Processing for Audio and Acoustics(WASPAA), pp. 19-22, 2001.

[6] G. Agostini, M. Longari and E. Pollastri, *Music instrument timbres classification with spectral features*, Proc. of ICME, pp. 97-102, 2001.

[7] T. Kitahara, M. Goto and H. G. Okuno, *Music instrument identification based on F0 dependent multivariate normal distribution*, Proc. of ICASSP, pp. 421-424, 2003.

[8] K. K. Paliwal and B. S. Atal, *Efficient vector quantization of LPC parameters at 24 bits/frame*, IEEE Transactions on Speech and Audio Processing, Vol. 1, pp. 3-14, Jan., 1993.