Musical Instrument Identification with Supervised Learning

Orchisama Das,
Center for Computer Research in Music and Acoustics
SUID : odas

Abstract—**In this paper, the classification of musical instruments using supervised learning is studied. A combined feature set including psychoacoustically relevant spectral features is used. A fairly small dataset consisting of 496 sound examples from 4 instruments - violin, clarinet, saxophone and bassoon is used. A very high classification accuracy is achieved on the test set with multi-class logistic regression and SVM with RBF kernel.**

I. Introduction

Musical instrument identification is an important and well-studied problem in the field of Music Information Retrieval. It is closely related to the perception of musical timbre [1]. Timbre is an ill-defined term, but it is thought to be related to the harmonic structure of sound in the frequency domain, which can be obtained by a Fourier transform. The timbre of an instrument is determined by its physical acoustics, the properties of the material it was made with, how the material has aged etc. Each instrument has a unique timbre. This is why two violins can sound very different from each other. Similarly, instruments in similar groups share timbre properties. For example, woodwind instruments can be grouped together by their timbre, which is very different from that of string instruments, since their sound production mechanisms are completely different.

Several attempts have been successfully made in identifying musical instruments with machine learning, such as [2], [3]. Musical instrument classification is important for automatic parsing of musical content on the web, and can act as a first-step in the more involved problem of music source separation. In [3], the author compares several features that are used in instrument identification. The selection of features is key in this problem, because it is closely tied to psychoacoustics, and we can hand engineer features that mimic human hearing. Based on the findings in [3], I have used a combined feature set with Mel-frequency cepstral coefficients, and warped linear prediction coefficients. Although I have used a small dataset comprising of only 4 instruments, I was able to achieve very high accuracy with a baseline logistic regression model and a support vector machine classifier with RBF kernel.

II. Dataset

I have used the **Bach 10** dataset [4] which consists of audio recordings of each part and the ensemble of ten pieces of four-part J.S. Bach chorales, as well as their MIDI scores. The audio recordings of the four parts (Soprano, Alto, Tenor and Bass) of each piece are performed by violin, clarinet, saxophone and bassoon, respectively. For each piece, the tracks are of the same length and aligned in time. I only used the solo instrument files, and separated each of 40 files into 2 s excerpts, giving a total of 496 files, i.e, 164 labeled sound files for each instrument. It is to be noted that these sound excerpts contain multiple note sequences at different pitches.

III. Features

A. Mel Frequency Cepstral Coefficients

It is a well known fact that human hearing is logarithmic in nature. In other words, humans have better frequency resolution at lower frequencies, and poor frequency resolution (but better temporal resolution) at higher frequencies. The Mel-scale divides the uniform frequency axis given by an FFT, into overlapping triangular filterbanks of different bandwidths, which is more psychoacoustically accurate. The relationship between Mel scale and linear frequency (f Hz) is given by:

$$m = 1125 \ln \left(1 + \frac{f}{700}\right) \tag{1}$$

Figure 1 shows a Mel filterbank with 12 bands. Mel frequency cepstral coefficients (MFCCs) are obtained by multiplying the frequency spectra with the Mel filterbank, summing the magnitudes in each band, taking their logarithm, and then computing the their Discrete Cosine Transform. The cepstral domain is ideal for observing harmonic structure, and the MFCCs give an indication of the spectral envelope, which is why they are an important feature used in speech recognition [5]. The log normalized Mel bank spectral energies of the 4 instruments used in the dataset are given in Fig 2.

B. Warped Linear Prediction Coefficients

Linear prediction is used in speech coding. In linear predictive coding, the signal is modeled as a weighted sum of p past samples, plus noise, where p is the LPC model order.

$$\begin{aligned} x(n) &= -\sum_{k=1}^{p} a_k x(n-k) + e(n) \\ X(z) &= \frac{E(z)}{1+\sum_{k=1}^{p} a_k z^{-k}} \end{aligned}$$

The a_k's are the LP coefficients, which can be obtained by the various methods, such as autocorrelation, covariance and Burg's method [6]. LPC gives an all-pole filter in the frequency (z) domain. The poles represent spectral peaks. For example, in speech the poles are located at speech formant frequencies.

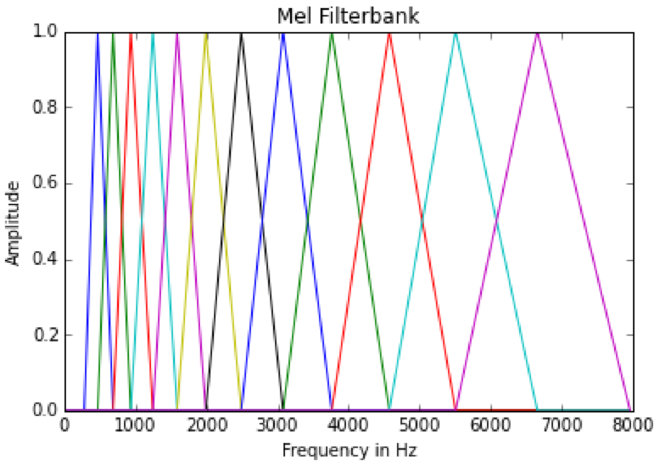

Fig. 1. 12 band Mel filterbank

Frequency warping [7] warps a uniform frequency axis to a non-uniform one. Warping is done by replacing the unit delay z^{-1} with a first-order allpass filter ζ^{-1}, given by:

$$\zeta^{-1} = \frac{z^{-1} - \rho}{1 - \rho z^{-1}} \tag{2}$$

where ρ determines the warping factor. By choosing $\rho = 0.7564$, we can map uniformly spaced points on the frequency axis to the non-uniform Bark scale [8], which is another log spaced psychoacoustic scale, like the Mel scale.

To get frequency warped LPCs,we simply replace the unit delays in (2) with ζ^{-1}, to get:

$$X_w(z) = \frac{E(z)}{1 + \sum_{k=1}^{p} a_k \zeta^{-k}} \tag{3}$$

This warps the pole locations, such that lower frequency peaks are narrower (in signal processing terms, they have a higher Q) and higher frequency peaks get more spread out. In essence, the warped LPCs (WLPCs) are very similar to the MFCCs, since they both represent psychoacoustically relevant spectral information. A comparison of the two spectra is given in Fig. 3

C. Feature Extraction

Features were computed using the **Essentia** audio signal processing library for Python [9]. Each sound file was broken into frames of length 1024 samples, with a hopSize of 512 samples. For each frame, two features were computed - MFCCs and warped LPCs. For MFCC calculation, the frames were windowed with a Hanning window. Then, their spectrum was calculated via an FFT, which was used to yield a 40 band Mel filterbank and 15 MFCCs per frame. Warped LPCs of order 15 were calculated directly on the time domain frame data using the autocorrelation method. The median values over all frames were selected as the feature set for both MFCCs and WLPCs. They were then standardized by subtracting the mean and normalizing by the standard deviation. The number of Mel Coefficients and order of LPC was selected based on what is generally used in speech processing literature.

IV. Classification Methods

A. Logistic regression

Multi-class logistic regression with L_1 regularization is used as a classification model. The conditional distribution of the ith output, $y^{(i)}$, having label k given the input data $x^{(i)}$ is given by :

$$P(y^{(i)} = k \mid x^{(i)}; \theta) = \frac{\exp(\theta_k^T x^{(i)})}{1 + \sum_{j=1}^{K-1} \exp(\theta_j^T x^{(i)})} \tag{4}$$

For n training examples and K classes, the likelihood function of logistic regression is :

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \left(\prod_{k=1}^{K} P(y^{(i)} = k \mid x^{(i)}; \theta)^{\mathbb{1}(y^{(i)}=k)} \right) \\ l(\theta) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}(y^{(i)} = k) \log P(y^{(i)} = k \mid x^{(i)}; \theta) \\ &+ \lambda ||\theta||_1 \end{aligned}$$

To get K optimum weight vectors θ_k, we maximize the log likelihood function with gradient descent methods. Once trained, we predict the class for the ith test example as:

$$\arg \max_{k=1,\ldots,K} P(y^{(i)} = k \mid x^{(i)}; \theta) \tag{5}$$

B. SVM with RBF kernel

A support vector machine (SVM) with an infinite dimensional feature map, i.e, the RBF kernel, is used as another classification model. Non-linear SVM tries to fit the maximum-margin hyperplane separating classes in a transformed feature space by solving the following constrained optimization problem :

$$\begin{aligned} &\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \\ &s.t\ y^{(i)}(w^T \phi(x^{(i)}) + b) \geq 1 - \zeta_i, \\ &\zeta_i \geq 0\ \forall\ i = 1, \ldots, n \end{aligned}$$

The dual of this optimization problem is:

$$\begin{aligned} &\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha + \mathbb{1}^T \alpha \\ &s.t\ y^T \alpha = 0, \\ &0 \leq \alpha_i \leq C\ \forall\ i = 1, \ldots, n \end{aligned}$$

Here, Q is a positive-semidefinite matrix whose i, j th element is $Q_{ij} = y^{(i)} y^{(j)} K(x^{(i)}, x^{(j)})$. $K(x^{(i)}, x^{(j)}) =< \phi(x^{(i)}), \phi(x^{(j)}) >$ is the RBF kernel given by:

$$K(x^{(i)}, x^{(j)}) = \exp\left(-\gamma ||x^{(i)} - x^{(j)}||^2\right) \tag{6}$$

If $x^{(i)}$ and $x^{(j)}$ are close to each other, then the Kernel function would approach a value of unity, otherwise, it is a small fractional number.

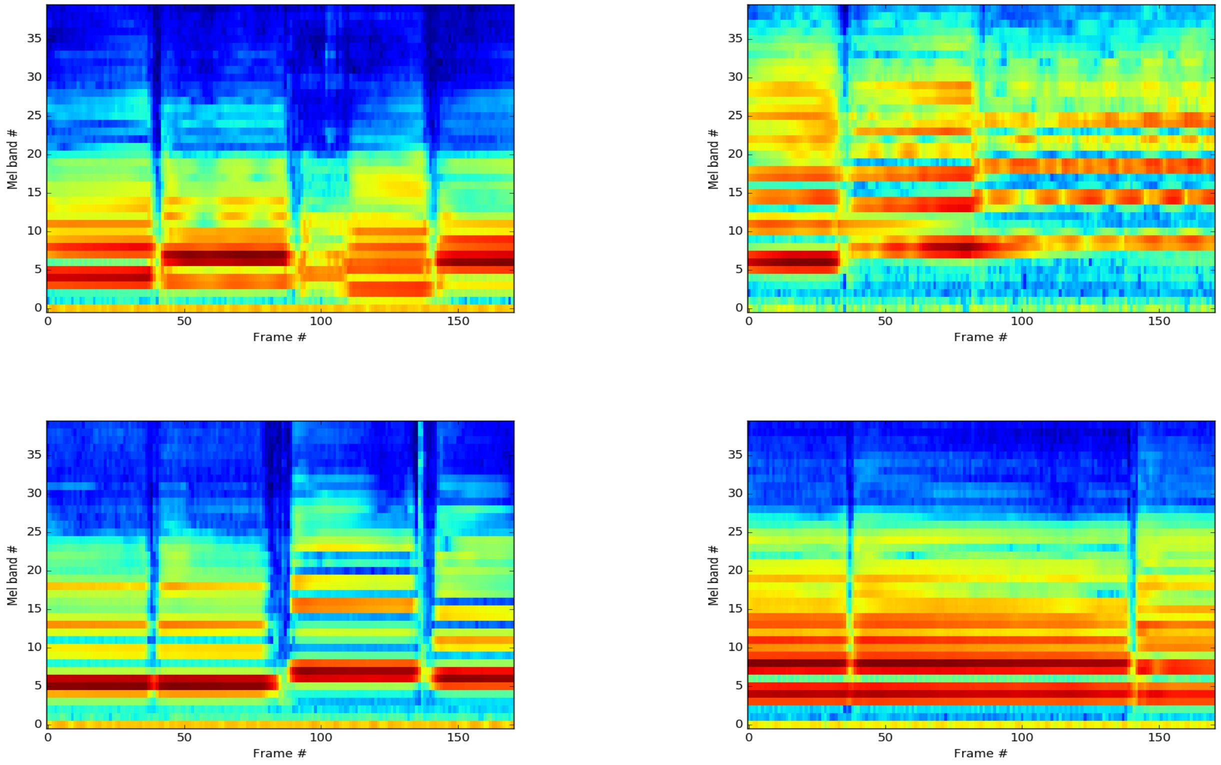

Fig. 2. Log normalized Mel bank spectral energies for bassoon, violin, clarinet and saxophone (top left to bottom right)

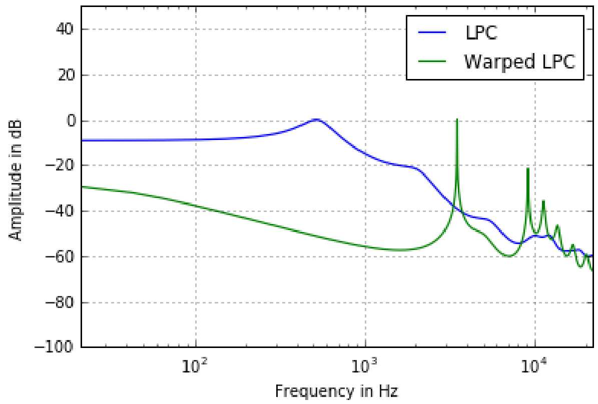

Fig. 3. LPC v/s Warped LPC frequency spectrum for order = 15

V. Experiments and Results

Out of 164 excerpts for each instrument, 75% were randomly assigned to the training set, and 25% to the test set. **Scikit-learn** [10] was used for training and classification.

A. Logistic Regression

1) Cross Validation: kFold cross-validation was performed on the training set, with $k = 3$ to choose the regularization parameter, $\lambda \in [10^{-3}, 10^{3}]$. $\lambda^* = 0.01$ was shown to give the best validation scores without overfitting. None of the 30 feature weights were dropped by L_1 regularization. The classification accuracy v/s convergence rate for the training set is given in Fig 4.

2) Test Results: 100% classification accuracy is achieved. The confusion matrix is given in Table I.

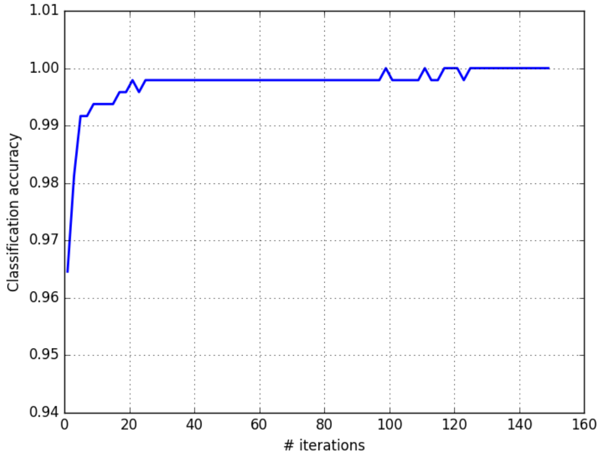

Fig. 4. Logistic regression training error

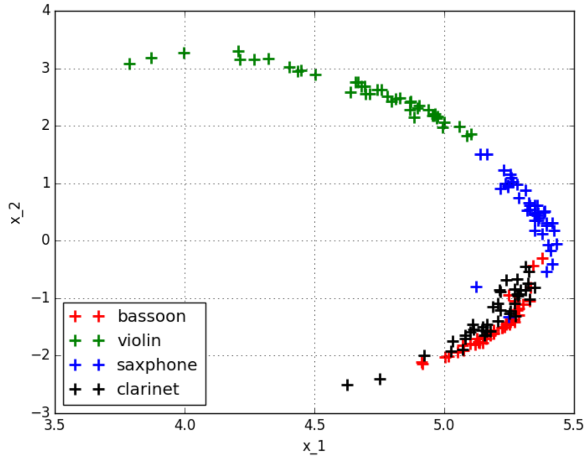

Fig. 5. Top 2 singular values of test set

TABLE I
CONFUSION MATRIX FOR LOGISTIC REGRESSION

	Bassoon	**Violin**	**Clarinet**	**Saxophone**
Bassoon	41	0	0	0
Violin	0	41	0	0
Clarinet	0	0	41	0
Saxophone	0	0	0	41

B. SVM with RBF kernel

1) Cross Validation: Again, kFold cross-validation was performed on the training set, with $k = 3$ to choose parameters $C \in [10^{-2}, 10^{2}]$, which is a regularization parameter that penalizes the width of the margin, and $\gamma \in [10^{-3}, 10^{2}]$, which is a parameter of the RBF kernel. With increasing γ, the Euclidean distance between the two elements in the kernel needs to reduce to have an effect on the value of the kernel. The optimal values selected after cross-validation are $C^{*} = 10, \gamma^{*} = 0.1$. After training, only 6 support vectors were found.

2) Test Results: 100% classification accuracy is achieved. The confusion matrix is given in Table II.

TABLE II
CONFUSION MATRIX FOR SVM WITH RBF KERNEL

	Bassoon	**Violin**	**Clarinet**	**Saxophone**
Bassoon	41	0	0	0
Violin	0	41	0	0
Clarinet	0	0	41	0
Saxophone	0	0	0	41

VI. DISCUSSION AND CONCLUSION

One reason why perfect accuracy is achieved could be because the spectra of these instruments is harmonic - there are no percussive attacks, which are stochastic in nature and harder to characterize with spectral features. MFCCs and WLPCs are adept at coding harmonic spectra. Moreover, the instruments are distinct from each other. Violin is a higher register string instrument with a rich spectrum, whereas bassoon is a lower register woodwind instrument that has fewer partial overtones. There is some possibility of confusion between saxophone and clarinet, since both are woodwind instruments with a single-reed mouthpiece, but that is avoided by these classification models.

To visualize the classification results, SVD was performed [11] on the test set features. The top 2 singular values were selected and plotted with their associated predicted labels in Fig. 5. The largest singular value explains 71% of the variance in the data. There is a significant overlap between the singular values of the bassoon and clarinet, but the other classes are distinct. Ultimately, we are only classifying 4 instruments - with more categories of instruments, classification accuracy is bound to decrease.

REFERENCES

[1] S. McAdams, "Musical timbre perception," *The psychology of music*, pp. 35–67, 2013.

[2] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 1, p. 943279, 2003.

[3] A. Eronen, "Comparison of features for musical instrument recognition," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 2001, pp. 19–22.

[4] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, 2010.

[5] C. Ittichaichareon, S. Suksri, and T. Yingthawornsuk, "Speech recognition using mfcc," in *International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July*, 2012, pp. 28–29.

[6] S. M. Kay, *Fundamentals of statistical signal processing: Practical algorithm development*. Pearson Education, 2013, vol. 3.

[7] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *Journal of the audio engineering society*, vol. 48, no. 11, pp. 1011–1031, 2000.

[8] J. O. Smith and J. S. Abel, "Bark and erb bilinear transforms," *IEEE Transactions on speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[9] D. Bogdanov and N. e. a. Wack, "Essentia: An audio analysis library for music information retrieval," in *14th Conference of the International Society for Music Information Retrieval (ISMIR)*. ISMIR, 2013, pp. 493–498.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[11] E. W. Weisstein, "Singular value decomposition," 2002.