MUSICAL NOTE AND INSTRUMENT CLASSIFICATION WITH LIKELIHOOD-FREQUENCY-TIME ANALYSIS AND SUPPORT VECTOR MACHINES

Mehmet Erdal Özbek[1], *Claude Delpha*[2], *and Pierre Duhamel*[2]

[1] Dept. of Electrical and Electronics Engineering
Izmir Institute of Technology, Urla, 35430, Izmir, Turkey
phone: + (90) 232 7506595, fax: + (90) 232 7506599
email: erdalozbek@iyte.edu.tr

[2] Laboratoire des Signaux et Systèmes (LSS)
Supélec, CNRS, Univ Paris Sud 11,
3, rue Joliot-Curie, 91192, Gif-sur-Yvette, France
phone: + (33) (0)1 69 85 17 64, fax: + (33) (0)1 69 85 17 65
emails: {claude.delpha, pierre.duhamel}@lss.supelec.fr

ABSTRACT

In this paper, we analyze the classification performance of a likelihood-frequency-time (LiFT) analysis designed for partial tracking and automatic transcription of music using support vector machines. The LiFT analysis is based on constant-Q filtering of signals with a filter-bank designed to filter 24 *quarter-tone frequencies of an octave. Using the LiFT information, features are extracted from the isolated note samples and classification of instruments and notes is performed with linear, polynomial and radial basis function kernels. Correct classification ratios are obtained for* 19 *instrument and* 36 *notes.*

1. INTRODUCTION

As automatic music transcription has been a popular research topic during the last years, many solutions are proposed to determine when and how long each instrument plays each note. The problem is very complex because of the possible combinations of instruments and notes. Partial solutions exist for both monophonic and polyphonic cases [1, 2]. However, there is not a complete solution for writing note symbols from the sound of played instruments [2].

The recognition or classification of musical instruments is one of the important steps in transcription and previous research on classification has concentrated on working with isolated notes. With the use of a sound sample collection which generally consist of isolated note samples of different instruments, the general classification problem is basically composed of calculating the features from the samples and classifying them with a learning algorithm [3]. Music information retrieval systems demand solutions for automatic classification of genre, composer, singer, song, or any other label which help to identify music on streams running especially over Internet. Therefore, new algorithms are now tested with not only isolated notes but also sound excerpts taken from commercial recordings.

There have been many attempts to solve the problem with different number of techniques which mainly decompose the problem into small problems and offer solutions for that specific part of the problem, for example determining the fundamental frequency, tempo, genre, timbre, etc. Using a wide range of techniques varying from speech processing research to more general signal processing techniques we now have a wide set of features [3, 4]. Features representing temporal, spectral and cepstral information are extracted independently or in a mixture. Feature extraction is followed by various classification algorithms including Gaussian Mixture Models and Support Vector Machines (SVM) which demonstrate successful classification rates [5].

As the inefficiency of using only temporal information or Fourier transform based spectral and cepstral information, time-frequency representations are required for music processing. By using constant time windows as in short-time Fourier transform (STFT), using variable length windows as a function of frequency as in constant-Q transform or scales in wavelets, the aim is to reveal the processes of music hidden in the time-frequency plane [6].

In this study, based on the constant-Q transform [7], we analyze the classification performance of likelihood-frequency-time analysis [8], designed for partial tracking and automatic transcription of music using support vector machines. Obtaining the likelihood-frequency-time information from the quarter-tone analysis of signals, feature vectors are extracted from the isolated note samples and used to classify the instruments and notes. The correct classification rates are evaluated.

The organization of the paper is as following: The likelihood-frequency-time analysis method is explained in the next section. In Section 3 a brief information on support vector machines is given. Section 4 will cover the simulation results for classification performance of support vector machine classifiers. Results are summarized and the future directions are discussed in Conclusion.

2. LIKELIHOOD-FREQUENCY-TIME ANALYSIS

The likelihood-frequency-time (LiFT) method [8] analyzes the output signal $y(n)$, considering the input signal as the sum of cosines

$$\begin{aligned} x_0(n) &= \sum_j a_0 \, cos(2\pi f_{0,j} n + \phi) \qquad (1) \\ &= \sum_j c_{0,j} \, cos(2\pi f_{0,j} n) + s_{0,j} \, sin(2\pi f_{0,j} n), \end{aligned}$$

and a white noise $b(n)$ where

$$y(n) = x_0(n) + b(n), \qquad (2)$$

with a Q-constant filter-bank composed of 24 filters whose center frequencies are set to quarter-tones. The main idea is to keep the same analysis structure of a signal for every octave while avoiding aliasing. Filters are designed as described in [7] with a quality factor $Q \approx 34$, which is highly selective.

Then, the time-frequency domain obtained from the filter-bank is analyzed statistically using a sliding window and a generalized likelihood approach is evaluated for each window by testing the two hypotheses whether there exist only noise in the output of the filter (H_0) or there exist both input signal and noise (H_1) . Under each of both hypotheses, the maximum probability density function for the values of cosine amplitude vector $\theta = (c_0\ s_0)^T$ is calculated and the generalized likelihood ratio is evaluated as

$$\Gamma = \frac{max_{\theta \in H_1} P_{H_1}}{max_{\theta \in H_1} P_{H_0}}. \tag{3}$$

Since Γ varies exponentially, the log-likelihood values are found using $\gamma = \log \Gamma$.

Although the LiFT analysis is designed both for time-domain where the samples of input signal are directly used and for frequency domain where the Fourier transform of the input signal is taken, in this study time-domain likelihood analysis is performed. Figure 1 shows an example of the likelihood-frequency-time plot of an input signal using the calculated log-likelihood values (γ) obtained for Alto Flute $A3$ note sample analyzed for 7 octaves. The likelihood values are normalized where the highest likelihood ratio value is shown as the darkest.

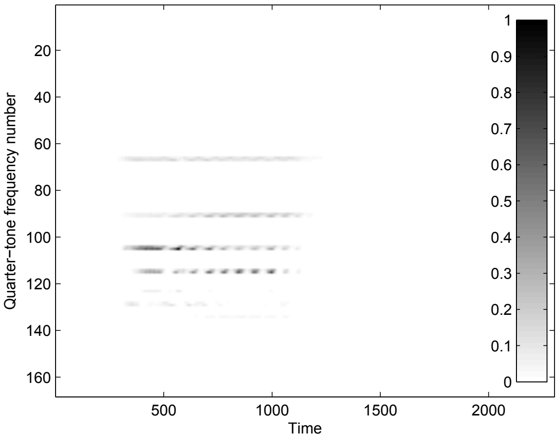

Figure 1: Likelihood-time-frequency plot of Alto Flute A3 note sample

Although in this work we only demonstrate results using monophonic samples, this approach is useful for polyphonic applications because of its ability to show multi-partials at the same time instants which may be extracted from the polyphonic instruments or any group of instruments playing simultaneously.

3. SUPPORT VECTOR MACHINES

The foundations of Support Vector Machines (SVM) have been developed by Vapnik based on statistical learning theory [9]. The theory which drove the initial development of SVM says that for a given learning task, with a given finite amount of training data, best generalization performance will be achieved when the capacity of the classification function is matched to the size of the training set [10]. The first application is introduced by Boser, Guyon and Vapnik [11] as a maximal margin classifier, with the training algorithm that automatically tunes the capacity of the classification function by maximizing the margin between the training patterns and the class decision boundary. When the training samples $\mathbf{x}$ of dimension n with the assigned labels y showing either of the two classes $(y_i \in \{-1,1\})$ are given, the algorithm searches for the optimal separating hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ so that

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0 \quad for\, \forall i, \tag{4}$$

under the constraint that the margin, given by $2/\|\mathbf{w}\|$ and defined as the distance between the hyperplane and the closest sample, is maximal. The training examples that are closest to the decision boundary which are usually a small subset of the training data form the resulting classification function and named as support vectors [12].

The maximum margin classifier is simple and proposed for problems which the patterns are linearly separable. However, when the data is not linearly separable or when the classes overlap because of noise, an additional cost function associated with misclassification is used [13]. Then a soft margin classifier is obtained by determining the trade-off between margin maximization and training error minimization. Nevertheless, when the patterns are not linearly separable one can still use the simple SVM or the soft margin classifier with a kernel function K, such that for all patterns in the input feature space X, (i.e. $\mathbf{x}, \mathbf{z} \in X$)

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z}), \tag{5}$$

where ϕ is a mapping from X to some higher (possibly infinite) dimensional feature space H where the patterns become linearly separable. The dot product in H can be computed without knowing the explicit form of ϕ using a substitution known as kernel trick. Then any function can be used to construct an optimal separating hyperplane in some feature space provided that Mercer's condition holds. Mercer's condition tells us whether or not a kernel is actually a dot product in some space [13].

The most common functions for kernels are the linear kernel

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z}), \tag{6}$$

polynomial kernel

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z} + 1)^d, \tag{7}$$

and radial basis function (RBF) kernel

$$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right). \tag{8}$$

There are also many kernels constructed for particular applications [14].

Although the SVM method is designed for two-class classification, multi-class classification is performed by the two common methods "one-vs-one" and "one-vs-all" (or one-vs-rest). Both consider the multi-class problem as a collection of two-class classification problems. For k-class classification, one-vs-all method constructs k classifiers where each classifier constructs a hyperplane between one class and the rest $k-1$ classes. A majority vote or some other measure is applied over the all possible pairs for decision. For the one-vs-one approach, $\frac{k(k-1)}{2}$ classifications are realized between each possible class pairs and similarly some voting scheme is applied for decision.

4. EXPERIMENTAL STUDY AND RESULTS

For this study, we use the University of Iowa Electronic Music Studios samples [15] of 19 mono recorded instruments (Flute, Alto Flute, Bass Flute, Oboe, $E\flat$ Clarinet, $B\flat$ Clarinet, Bass Clarinet, Bassoon, Soprano Saxophone, Alto Saxophone, French Horn, $B\flat$ Trumpet, Tenor Trombone, Bass Trombone, Tuba, Violin, Viola, Cello, Double Bass). The group of notes in this library sampled at 44100 Hz are separated and labeled according to each instrument and note. The dynamic ranges fortissimo (ff), mezzo forte (mf), and pianissimo (pp) are all included with or without vibrato depending on the instrument and for string instruments played with bowing (arco) and plucking (pizzicato), making a database with a total of nearly 5000 samples. Then the LiFT analysis is performed for 7 octaves for each of these note samples. Likelihood values of $7 \times 24 = 168$ quarter-tone frequencies are calculated. The feature vectors are extracted from these likelihood values and used for instrument and note classification. Various normalization schemes were tested and their effect on the classification performance was investigated. For example the features are standardized to have zero mean and unit variance with $\hat{x} = (x - \mu_x)/\sigma_x$ where μ_x and σ_x are the mean and the standard deviation of each feature x. However, the normalization of feature vectors to be in $[0,1]$ is found to give the best performance, therefore all feature vectors are normalized accordingly for the results presented here.

Support vector machines with linear, polynomial and radial basis function (RBF) kernels are used. Parameters of polynomial kernel and RBF kernel are also varied. One-vs-all approach is chosen for multi-class classification. The half of the features for each class are used for training and remaining half is left for testing. Correct classification ratios are obtained as the percentage of correctly classified class to the number of class samples. Results are the mean values of 10 different realizations.

4.1 Instrument Classification

For instrument classification of 19 instruments a feature vector is selected in two steps. In the first step, the maximum value of likelihood for each note sample is selected as a feature vector. This is a very simple vector and does not include and express the time information of the samples because it only takes information along the quarter-tone frequency number. Then as a second step, time information is included by selecting 10 time instants equally taken according to the length of the note sample and calculating the maximum value of likelihood for each time instant. Thus the feature vector for step 2 is not a vector composed of only showing likelihood values for all the duration of note sample (168×1) but a vector showing the likelihood values for 10 time instants (1680×1).

Figure 2 shows the best performance results for polynomial kernel obtained with $d = 2$, where the second step increases the performance slightly. This is also valid for the RBF kernel as given in Figure 3. Therefore throughout the paper, results obtained with second step are used.

As it is seen from the results that Bass has the highest correct classification results due to its frequency range. However the selection of different kernels or parameters does not have a major effect on classification. Also notice that this is a multi-class classification performed with 19 instruments. Any subclassification or grouping will possibly increase the correct classification rates.

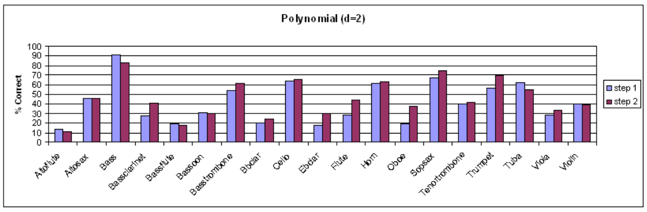

Figure 2: Classification of 19 instruments with polynomial kernel

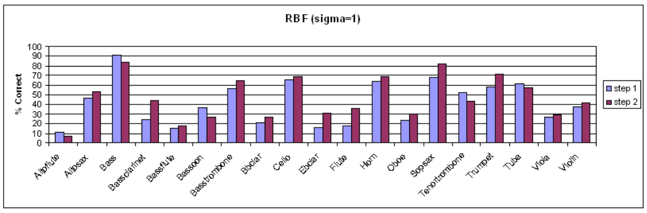

Figure 3: Classification of 19 instruments with RBF kernel

For example in [16], spectral features composed of 18 descriptors are extracted and the recognition of individual instruments having 17, 20 and 27 instrument samples are done with different classification techniques including SVM. RBF kernel is found to give the best results. Although the family of saxophones are combined in a single instrument class, an error rate of 19.8% in the classification of 17 instruments is achieved. Table 1 shows the classification performance with SVM given in [16]. It is obvious that increasing the number of instruments decrease the success rates. Nevertheless, a subclassification based on instrument family or pizzicati/sustained grouping increase the correct classification rates.

Table 1: Success rates for different instruments in [16]

	Success rate (%)
17 instruments	80.2
20 instruments	78.5
27 instruments	69.7
27 instr. family discrimination	77.6
27 instr. pizz./sust. discrimination	88.7

Therefore a small subset of the instruments is selected as the 5 woodwind instruments (Alto Saxophone, Bassoon, $B\flat$ Clarinet, Flute, Oboe). The correct classification results given with bold font on Table 2 demonstrate the performance using linear, polynomial and RBF kernels. Best result of RBF kernel is obtained when $\sigma = 1$. The results of the work in [5] are given for comparison. Better performance for $B\flat$ Clarinet is achieved. Note that in [5] polynomial kernel with $d = 5$ and RBF kernel results were not available.

The performance results of 19 instrument classification are compared with 5 instrument classification in Table 3. The ratios of only 5 instruments are shown with bold font. Obviously, for every kernel and its parameter the ratios of 5 instrument case are higher than the 19 instrument case. While the best average results for 19 instruments without normalization is 46.6% with RBF kernel $\sigma = 1$, the best average of these specific 5 instruments among 19 is 34.7%. However, the mean value obtained for only 5 instrument case is 68.1%.

Table 2: Classification of 5 woodwind instruments and comparison with the work in [5]

% correct	Alto Sax	Bassoon	$B\flat$ Clarinet	Flute	Oboe
Linear	**66.6** 73.4	**82.4** 88.0	**45.9** 31.2	**69.2** 82.8	**70.2** 66.9
Polynomial (d=2)	**72.1** 69.2	**75.1** 88.0	**40.6** 33.0	**72.9** 76.3	**68.7** 66.4
Polynomial (d=3)	**68.8** 69.9	**73.6** 87.2	**36.4** 27.0	**76.9** 86.8	**63.5** 74.8
Polynomial (d=4)	**64.2** 69.0	**71.2** 87.6	**35.2** 28.5	**80.1** 86.4	**59.4** 75.9
Polynomial (d=5)	**59.8** -	**67.1** -	**32.3** -	**81.2** -	**55.8** -
RBF (σ=1)	**77.2** -	**76.4** -	**41.2** -	**72.4** -	**73.3** -

Selecting a small subset corresponds to an almost double increase in the correct classification ratio.

Table 3: Comparison of the classifications using 19 and 5 instruments

% correct	Alto Sax	Bassoon	$B\flat$ Clarinet	Flute	Oboe
Linear	31.3 **66.6**	26.9 **82.4**	19.7 **45.9**	41.5 **69.2**	34.5 **70.2**
Polynomial (d=2)	45.6 **72.1**	29.6 **75.1**	24.5 **40.6**	44.4 **72.9**	37.4 **68.7**
Polynomial (d=3)	43.8 **68.8**	25.2 **73.6**	23.2 **36.4**	39.4 **76.9**	34.7 **63.5**
Polynomial (d=4)	41.0 **64.2**	17.5 **71.2**	23.0 **35.2**	35.0 **80.1**	33.7 **59.4**
Polynomial (d=5)	38.4 **59.8**	14.0 **67.1**	21.5 **32.3**	29.8 **81.2**	28.4 **55.8**
RBF (σ=1)	53.8 **77.2**	27.4 **76.4**	26.1 **41.2**	36.3 **72.4**	30.0 **73.3**

4.2 Note Classification

Remember that the LiFT analysis is mainly designed for partial tracking, it is more likely that correct classification performance will increase in the classification of notes. As in instrument classification when the number of classes is high, it is difficult to obtain a high correct classification ratio. Nevertheless, the classification of a single note among all possible notes is important hence all database (except piano) note samples need to be used. However, because of the lack of samples available for each note, three octave range from $C3$ to $C6$ is selected where these 36 notes are in the common range for most of the instruments. As the number of note samples per instrument is not the same, the number of training and test samples vary. However, for each class at least 50 samples are taken for the accuracy of the classification results with a total of nearly 3000 samples. To our knowledge this is the first trial of a note classification using such number of notes.

Figure 4 shows the performance results for both steps. Results with polynomial kernel with parameters greater than (d=2) are not shown for the clarity of figures and because their performance are not better with respect to their nonlinearity expected to discriminate better. Both figures demonstrate that correct classification ratios over 40% and even 50% (for step 1 except linear kernel, which is lower because of the simple feature and kernel function) are achieved. As the number of available notes between $C3$ and $C4$ is more than the interval $C4$-$C5$ or interval $C5$-$C6$, the average correct classification ratio for that octave is higher. With a large sample database it is expected to have higher ratios. Also, even the ratios do not exceed 80% it is very likely that using a subclassification will increase the correct classification ratios. For example, the notes of string instruments played by plucking are removed from the note database and classifications are performed. Results obtained by using step 2 are given in Figure 5.

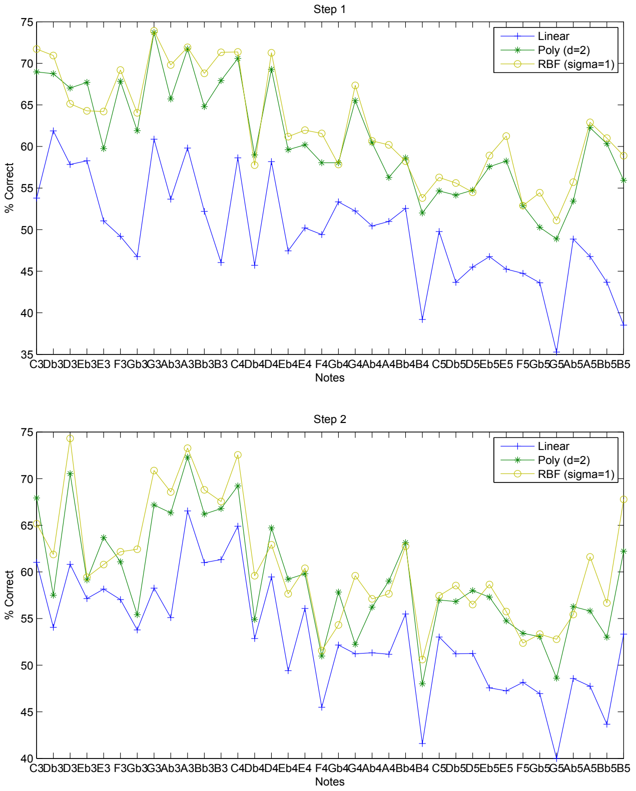

Figure 4: Classification of notes from $C3$ to $C6$ with both feature sets

The best average results for 36 notes without normalization is found as 62.6% in step 1 and 60.8% in step 2 with RBF kernel $\sigma = 1$. With the removal of the notes played by plucking, the best average results of these notes is calculated as 68.9% for step 2. Therefore even with less samples, selecting a better subset corresponds to a 8% increase in the correct classification ratio.

Moreover, with a pre-classifier which aims to find the octave number, the number of classes will be limited to 12 and better classification could be achieved. Notice that the time information which is included with the second step is not effective in note classification due to the discriminative power of frequency patterns over the notes extracted by the quarter-tone filtering of LiFT analysis.

5. CONCLUSION

In this paper, likelihood-frequency-time information is used for classification of instruments and notes. With a Q-constant filter-bank composed of 24 filters whose center frequencies are set to quarter-tones, the time-frequency domain is analyzed by testing the two hypotheses with the generalized likelihood ratio. For each isolated note sample, feature vectors are extracted from the likelihood values. Although the

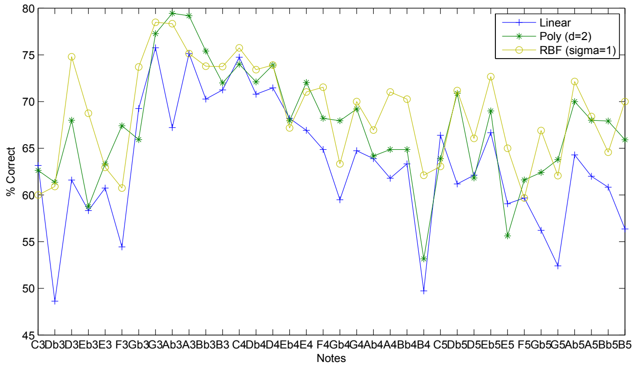

Figure 5: Classification of notes from $C3$ to $C6$ without plucked string samples

number of available samples effected the correct classification performance directly, multi-class classification of 19 instruments and 36 notes is performed with support vector machines using linear, polynomial and RBF kernels with varying parameters, and correct classification ratios are obtained. For most of the instruments and notes, the best performance is obtained using RBF kernel with parameter $\sigma = 1$.

For instrument classification, the performance of SVM with the features extracted from LiFT analysis is evaluated using two different works having large and small number of instruments. For large number of instruments, it is shown that the correct classification ratios tend to decrease because of the high number of classes in multi-class classification. The closeness among the classes increase the misclassifications resulting an overall decrease in performance. Therefore an additional classification based on family or any other grouping is expected to be effective. This is demonstrated with the small number of instrument case, although the classification among a family seem to be more difficult, better classification ratios than the large number of instruments are achieved. For the same small subset of woodwind instruments better classification ratio for $B\flat$ Clarinet is observed.

For note classification, the classification of 36 notes with more than 3000 note samples is novelly performed and correct classification ratios are obtained. The selection of a better subset of notes such as choosing the note samples of string instruments played with bowing, is shown to give better correct classification ratios even with the less number of samples.

The LiFT analysis is found to be more adequate for note classification than instrument classification because of the quarter-tone filtering extracting the partials. Besides, the time information of samples is not fully represented in the feature vectors. The proper selection of the discriminating features from LiFT will definitely help to achieve better classification performance. In the future work, the pre-classifier for octave selection will be included in the system to obtain better correct classification ratios. A decomposition of the problem to its minimum level may also be another issue.

REFERENCES

[1] I. Bruno and P. Nesi, "Automatic Music Transcription Supporting Different Instruments", *Journal of New Music Research*, vol. 34, no. 2, pp. 139–149, 2005.

[2] A. P. Klapuri, "Automatic Music Transcription as We Know it Today", *Journal of New Music Research*, vol. 33, no. 3, pp. 269–282, 2004.

[3] P. Herrera-Boyer, G. Peeters, and S. Dubnov, "Automatic Classification of Musical Instrument Sounds", *Journal of New Music Research*, vol. 32, no. 1, pp. 3–21, 2003.

[4] S. Essid, G. Richard, and B. David, "Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies", *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, vo. 1, pp. 68–80, 2006.

[5] S. Essid, G. Richard, and B. David, "Musical Instrument Recognition on Solo Performances", in *EUSIPCO'04*, Vienna, Austria, September 7-10, 2004.

[6] W. J. Pielemeier, G. H. Wakefield, and M. H. Simoni, "Time-Frequency Analysis of Musical Signals", *Proc. of IEEE*, vol. 84, no. 9, pp. 1216–1230, 1996.

[7] J. C. Brown, "Calculation of a Constant-Q Spectral Transform", *Journal of Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[8] V. Verfaille, P. Duhamel, and M. Charbit, "LIFT: Likelihood-Frequency-Time Analysis for Partial Tracking and Automatic Transcription of Music", in *DAFX-01*, Limerick, Ireland, December 6–8, 2001.

[9] V. N. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.

[10] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", *Knowledge Discovery and Data Mining*, vol. 2, no. 2, pp. 121–167, 1998.

[11] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers ", in *Proc. 5th Annual Workshop on Computational Learning Theory*, Pittsburgh, USA, July 1992, pp. 144–152.

[12] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.

[13] C. Cortes and V. Vapnik, "Support Vector Networks", *Machine Learning*, vol. 20, pp. 273–297, 1995.

[14] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[15] The University of Iowa Electronic Music Studios. http://theremin.music.uiowa.edu

[16] G. Agostini, M. Longari, and E. Pollastri, "Musical Instrument Timbres Classification with Spectral Features", *EURASIP Journal on Applied Signal Processing*, no. 1, pp. 5–14, 2003.