MUSICAL INSTRUMENT RECOGNITION USING BIOLOGICALLY INSPIRED FILTERING OF TEMPORAL DICTIONARY ATOMS

Steven K. Tjoa and K. J. Ray Liu
Signals and Information Group, Department of Electrical and Computer Engineering
University of Maryland – College Park, MD 20742 USA
{kiemyang, kjrliu}@umd.edu

ABSTRACT

Most musical instrument recognition systems rely entirely upon spectral information instead of temporal information. In this paper, we test the hypothesis that temporal information can improve upon the accuracy achievable by the state of the art in instrument recognition. Unlike existing temporal classification methods which use traditional features such as temporal moments, we extract novel features from temporal atoms generated by nonnegative matrix factorization by using a *multiresolution gamma filterbank*. Among isolated sounds taken from twenty-four instrument classes, the proposed system can achieve 92.3% accuracy, thus improving upon the state of the art.

1. INTRODUCTION

Advances in sparse coding and dictionary learning have influenced much of the recent progress in musical instrument recognition. Many of these methods depend upon nonnegative matrix factorization (NMF) – a popular, convenient, and effective method for decomposing matrices – to obtain low-rank approximations of audio spectrograms [9]. NMF yields a set of vectors, spectral atoms, which approximately span the frequency space of the spectrogram, and another set of vectors, temporal atoms, which correspond to the temporal activation of each spectral atom. The spectral atoms can then be classified by instrument using features such as mel-frequency cepstral coefficients (MFCCs).

While these methods are effective in exploiting the spectral redundancy in a signal, redundancy remains in the *temporal* domain. Psychoacoustic studies have shown that spectral and temporal information are equally important in the definition of acoustic timbre [10]. Classification methods that only utilize spectral information are discarding the potentially useful temporal information that could be used to improve classification performance.

In this paper, we combine advances in dictionary learning, auditory modeling, and music information retrieval to propose a new timbral representation. This representation is inspired by another widely accepted timbral model, the cortical representation, which estimates the spectral and temporal modulation content of the auditory spectrogram. Our method of extracting temporal information uses a *multiresolution gamma filterbank* which is computed from the temporal atoms extracted from spectrograms using NMF. Extracting and classifying this feature is *simple yet effective* for musical instrument recognition.

After defining the proposed feature extraction and classification method, we test the hypothesis that the proposed feature improves upon the accuracy achievable by the state of the art in musical instrument recognition. For isolated sounds, we show that temporal information can be used to build a classifier capable of 72.9% accuracy when tested among 24 instrument classes. However, when combining temporal and spectral features, the proposed classifier can achieve an accuracy of **92.3%**, thus reflecting state of the art performance.

2. TEMPORAL INFORMATION

Temporal information is incorporated into timbral models in different ways. Many attempts to incorporate temporal information use features such as the temporal centroid, spread, skewness, kurtosis, attack time, decay time, slope, and locations of maxima and minima [5,6]. One timbral representation, the *cortical representation*, incorporates both spectral and temporal information. Essentially, the cortical representation embodies the output of cortical cells as sound is processed by earlier stages in the auditory system. Fig. 1 illustrates the relationship between the early and middle stages of processing in the mammalian auditory system. The early stage models the transformation by the cochlea of an acoustic input signal into a neural representation known as the auditory spectrogram, while the middle stage models the analysis of the auditory spectrogram by the primary auditory cortex.

One property of cortical cells, the spectrotemporal receptive field (STRF), summarizes the way a single cortical cell responds to a stimulus. Mathematically, the STRF is like a two-dimensional impulse response defined across time and frequency. Each STRF has three parameters: scale, rate, and orientation. Scale defines the spectral resolution of an STRF, rate defines its temporal resolution, and orientation determines if the STRF selects upward or down-

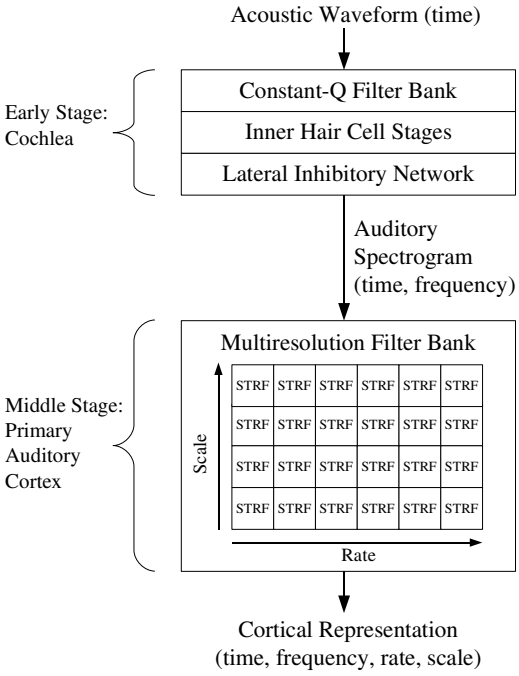

Figure 1. Early and middle stages of the auditory system. The auditory spectrogram is convolved across time and frequency with STRFs of different rates and scales to produce the four-dimensional cortical representation.

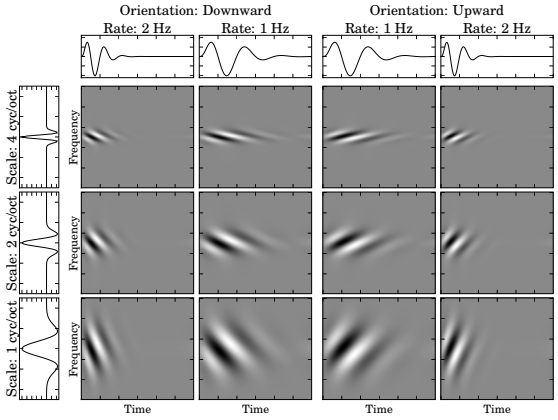

Figure 2. Twelve example STRFs. Together, they constitute a filterbank. The left six STRFs select downward-modulating frequencies, and the right six STRFs select upward-modulating frequencies. Top row: seed functions for rate determination. Left column: seed functions for scale determination.

ward frequency modulations. Fig. 2 illustrates the STRF as a function of these three parameters. Each cortical cell can be interpreted as a filter whose impulse response is an STRF with a particular rate, scale, and orientation. Therefore, a collection of cortical cells constitutes a filterbank. Indeed, it turns out that the cortical representation is mathematically equivalent to a multiresolution wavelet filterbank [2].

Despite the biological relevance between the cortical representation and timbre, this representation has disadvantages for classification purposes. First, because the cortical representation is a complex-valued four-dimensional filterbank output, it is massively redundant. Like many types of redundant data, the cortical representation could benefit from some form of coding, decomposition, or dimensionality reduction. However, proper application of these tools to the cortical representation for engineering purposes such as speech recognition and MIR is not yet well understood. Therefore, these are ongoing areas of research [11]. Second, the STRF is not time-frequency separable [2]. In other words, computation of the cortical representation cannot be decomposed into two procedures that operate on the time and frequency dimensions separately. Because spectral and temporal information require different classification methods, this obstacle impedes classification.

Unlike the cortical representation, the spectrogram computed via short-time Fourier transform (STFT) is easily decomposed, particularly for musical signals. For example, many works have applied decomposition methods to magnitude spectrograms of musical sounds in order to identify a set of spectral and temporal basis vectors from which the magnitude spectrogram can be parameterized [15]. One such decomposition method is NMF [9]. Given an element-wise nonnegative matrix $\mathbf{X}$, NMF attempts to find two nonnegative matrices, $\mathbf{A}$ and $\mathbf{S}$, that minimize some divergence between $\mathbf{X}$ and $\mathbf{AS}$. Among the algorithms that can perform this minimization, one of the most convenient algorithms uses a multiplicative update rule during each iteration in order to maintain nonnegativity of the matrices $\mathbf{A}$ and $\mathbf{S}$ [9].

Many researchers have already demonstrated the usefulness of NMF for separating a musical signal into individual notes [7,15,16]. By first expressing a time-frequency representation of the signal as a matrix, these methods decompose the matrix into a summation of a few individual *atoms*, each corresponding to one musical source or one note. Fig. 3 illustrates the use of NMF upon the spectrogram of a musical signal. We define each column of $\mathbf{A}$ as a *spectral atom* and each row of $\mathbf{S}$ as a *temporal atom*. The temporal atoms usually resemble envelopes of known sounds, particularly in musical signals. For example, observe the difference between the profiles of the temporal atoms in Fig. 3. The three beats generated by the kick drum share the same temporal profiles, and the two beats generated by the snare drum share the same profiles. This general observation motivates the hypothesis that the energy distribution of temporal NMF atoms is a valid timbral representation that can be used to classify instruments.

In the next section, we propose one technique that extracts timbral information from temporal NMF atoms similar to that of the cortical representation. Our technique uses a *multiresolution gamma filterbank* to perform multiresolution analysis upon the factorized spectrogram. However, unlike the cortical representation, this multiresolution analysis is particularly suited to the energy profiles contained in the temporal NMF atoms.

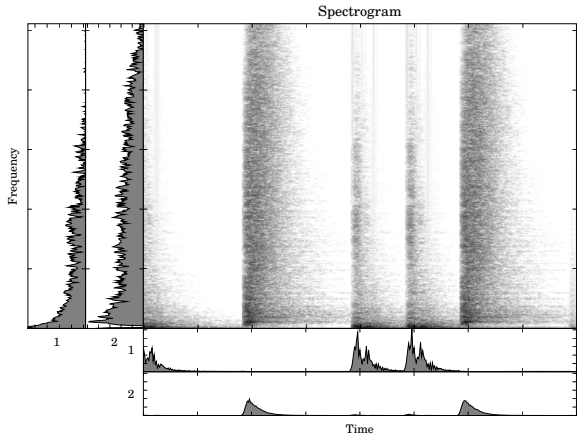

Figure 3. The NMF of a spectrogram drum beats. Component 1: kick drum. Component 2: snare drum. Top right: $\mathbf{X}$. Left: $\mathbf{A}$. Bottom: $\mathbf{S}$.

3. PROPOSED METHOD: MULTIRESOLUTION GAMMA FILTERBANK

The multiresolution gamma filterbank is a collection of gamma filters. For this work, we define the gamma kernel to be

$$g(t; n, b) = \alpha t^{n-1} e^{-bt} u(t) \tag{1}$$

where $b > 0$, $n \geq 1$, $u(t)$ is the unit step function, and

$$\alpha = \sqrt{\frac{(2b)^{2n-1}}{\Gamma(2n-1)}} \tag{2}$$

ensures that $\int |g(t; n, b)|^2 dt = 1$ for any value of n and b, where $\Gamma(n)$ is the Gamma function. Let I be the total number of gamma filters in the filterbank. For each $i \in \{1, ..., I\}$, define the correlation kernel (i.e., time-reversed impulse response) of each gamma filter to be

$$g_i(t) = g(t; n_i, b_i). \tag{3}$$

The set of kernels $\{g_1, g_2, ..., g_I\}$ defines the *multiresolution gamma filterbank*. Fig. 4 illustrates some example kernels of the filterbank.

For each i, let the filter output be the cross-correlation between the input atom, $s(t)$, and the kernel, $g_i(t)$:

$$y_i(\tau) = \int_{-\infty}^{\infty} s(t) g_i(t - \tau) dt \tag{4}$$

The set of outputs $\{y_1, y_2, ..., y_I\}$ from the filterbank is called the *multiresolution gamma filterbank response* (MGFR).

The gamma filter has convenient temporal properties. We define the *attack time* of the kernel $g(t)$ to be the time elapsed until the kernel achieves its maximum. By differentiating $\log g(t)$, we determine the attack time to be

$$t_a = (n - 1)/b \text{ seconds.} \tag{5}$$

Fig. 4 illustrates the relationship between the attack time and the parameter b. Also, as t becomes large, $\log g(t) \approx$

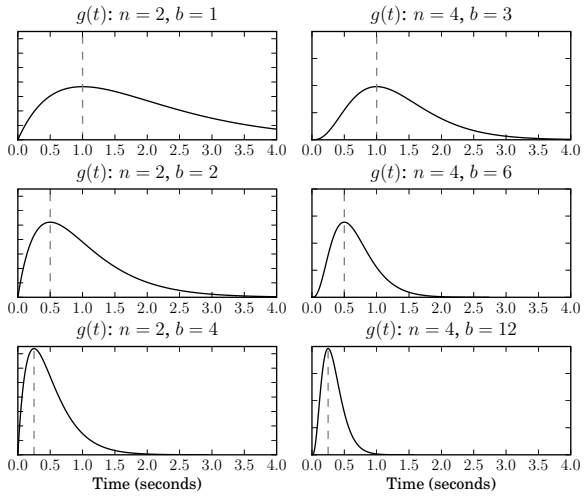

Figure 4. Kernels of gamma filters. The dashed vertical line indicates the location of the maxima. Left column: $n = 2$. Right column: $n = 4$.

$-bt$ plus a constant. Therefore, b is the decay parameter of $g(t)$, where we define the *decay rate* of $g(t)$ to be

$$r_d = 20b \log_{10} e \approx 8.7b \text{ dB per second.} \tag{6}$$

Together, these two temporal properties imply that a gamma kernel with *any* attack time and decay rate can be created from the proper combination of n and b.

Fig. 5 illustrates the operation of the multiresolution gamma filterbank. When a temporal NMF atom is sent through the multiresolution gamma filterbank, the MGFR reveals the strength of the attacks and decays of the atom's envelope for different values for n and b. Observe how the filterbank response is largest for those filters whose attack time matches that of the input atom.

The multiresolution gamma filterbank behaves like a set of STRFs. Both systems perform multiresolution analysis on the input data. Each STRF passes a different spectrotemporal pattern depending upon the rate and scale. In fact, the seed function used to determine the rate of an STRF is a gammatone kernel – a sinusoid whose envelope is a gamma kernel. By altering the parameters of the gammatone kernel, STRFs can select different rates. Similarly, in the multiresolution gamma filterbank, each filter passes different envelope shapes depending upon the parameters n and b which completely characterize the attack and decay of the envelope. Intuitively, the filter with kernel $g_i(t)$ passes envelopes with attack times equal to $(n_i - 1)/b_i$ seconds and envelopes with decay rates equal to $8.7b_i$ dB per second.

4. PROPOSED FEATURE EXTRACTION AND CLASSIFICATION

To extract a shift-invariant feature from the MGFR, we compute the norm for each filter response:

$$z_i = \left(\int_{-\infty}^{\infty} |y_i(t)|^p dt \right)^{1/p} \tag{7}$$

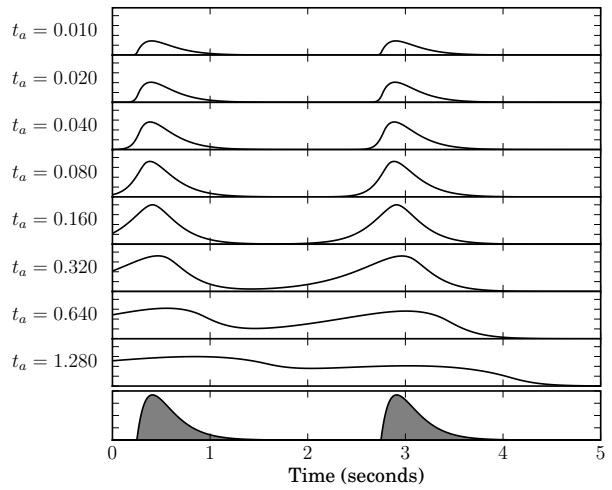

Figure 5. Top: MGFR as a function of time for $n = 2$. Bottom: input atom containing two pulses with attack times of 160 ms.

The vector $\mathbf{z} = [z_1, z_2, ..., z_I]$ is the extracted feature vector. To eliminate scaling ambiguities among the input atoms, every feature vector $\mathbf{z}$ is normalized to have unit Euclidean norm. Different choices of p provide different interpretations of $\mathbf{z}$. For this work, we use $p = \infty$. Our future work will include an investigation into the impact of p on classification performance.

The *proposed feature extraction algorithm* is summarized below.

1. Perform NMF on the magnitude spectrogram, $\mathbf{X}$, to obtain $\mathbf{A}$ and $\mathbf{S}$.
2. Initialize the multiresolution gamma filterbank in (3).
3. For each temporal atom (i.e., row of $\mathbf{S}$), compute the MGFR in (4).
4. Compute the feature vector $\mathbf{z}$ in (7).

Finally, we formulate the instrument recognition problem as a typical supervised classification problem: given a set of training features extracted from signals of known musical instruments, identify all of the instruments present in a test signal. To perform supervised classification, temporal atoms are extracted from training signals of known musical instruments using NMF. The feature vector $\mathbf{z}$ computed from the atom plus its instrument label are used for training. To predict the label of an unknown sample, $\mathbf{z}$ is extracted from the unknown sample and classified using the trained model.

An advantage of the proposed feature extraction and classification procedure is its *simplicity*. The proposed system requires no rule-based preprocessing. Unlike other systems that contain safeguards, thresholds, and hierarchies, the proposed system uses straightforward filtering and a flat classifier. As the next section shows, this simple procedure can achieve state-of-the-art accuracy for instrument recognition.

n	b	t_a	n	b	t_a
1.2	0.200	1.000	1.5	0.500	1.000
1.2	0.250	0.800	1.5	0.625	0.800
1.2	0.333	0.600	1.5	0.833	0.600
1.2	0.500	0.400	1.5	1.25	0.400
1.2	1.00	0.200	1.5	2.50	0.200
1.2	2.00	0.100	1.5	5.00	0.100
1.2	4.00	0.050	1.5	10.0	0.050
1.2	10.0	0.020	1.5	25.0	0.020
2.0	1.00	1.000	3.0	2.00	1.000
2.0	1.25	0.800	3.0	2.50	0.800
2.0	1.67	0.600	3.0	3.33	0.600
2.0	2.50	0.400	3.0	5.00	0.400
2.0	5.00	0.200	3.0	10.0	0.200
2.0	10.0	0.100	3.0	20.0	0.100
2.0	20.0	0.050	3.0	40.0	0.050
2.0	50.0	0.020	3.0	100	0.020

Table 1. Gamma filterbank parameters used in the following experiments.

5. EXPERIMENTS

We perform experiments on an extensive set of isolated sounds. The data set for these experiments combines samples from the University of Iowa database of Musical Instrument Samples [4], McGill University Master Samples [14], the OLPC Samples Collection [13], and the Freesound Project [12]. All of these samples consist of isolated sounds generated by real musical instruments. We have parsed the audio files such that each file consists of a single musical note (for harmonic sounds) or beat (for percussive sounds).

From each input signal, $x(t)$, we obtain the magnitude spectrogram, $\mathbf{X}$, via STFT using frames of length 46.4 ms (i.e., 2048/44100) windowed using a Hamming window and a hop size of 10.0 ms. Then, we perform NMF using the Kullback-Leibler update rules [9] with an inner dimension of $K = 1$ to obtain $\mathbf{A}$ and $\mathbf{S}$. When applicable, we use a multiresolution gamma filterbank of thirty-two filters with the parameters shown in Table 1. These attack times and decay rates cover a wide range of sounds produced by common musical instruments. Each 32-dimensional feature vector, $\mathbf{z}$, is then classified.

For supervised classification, we use the LIBSVM implementation [1] of the support vector machine (SVM) with the radial basis kernel. For multiple classes, LIBSVM uses the one-versus-one classification strategy by default. The remaining programs and simulations were written entirely in Python using the SciPy package [8]. Source code is available upon request.

In total, there are 3907 feature vectors collected among twenty-four instrument classes. Table 2 summarizes this data set. With few exceptions [3], this selection of instruments is more comprehensive than any existing work on isolated instrument recognition. Recognition accuracy for class c is defined to be the percentage of the feature vectors whose true class is c that are correctly classified by the SVM as belonging in class c. Overall recognition accuracy is the average of the accuracy rates for each class.

Instrument	#	S	T	ST
Bassoon	131	**99.2**	75.6	96.9
Clarinet	145	80.7	73.1	**86.2**
Flute	236	84.7	60.6	**89.0**
Oboe	118	72.0	77.1	**91.5**
Saxophone	196	**93.4**	65.8	86.7
Horn	92	80.4	62.0	**85.9**
Trombone	99	**93.9**	53.5	89.9
Trumpet	236	97.5	82.2	**97.9**
Tuba	111	98.2	75.7	**99.1**
Cello	349	94.8	89.7	**97.4**
Viola	309	**94.2**	67.6	90.9
Violin	390	**97.2**	86.2	96.2
Cello Pizz.	321	98.1	87.5	**98.4**
Viola Pizz.	254	**99.6**	81.9	**99.6**
Violin Pizz.	315	97.5	85.4	**99.0**
Glockensp.	10	**100.0**	90.0	**100.0**
Guitar	27	51.9	29.6	**63.0**
Marimba	39	46.2	25.6	**79.5**
Piano	260	95.0	89.2	**98.5**
Xylophone	13	61.5	53.8	**84.6**
Kick	90	98.9	95.6	**100.0**
Snare	86	96.5	88.4	**98.8**
Timpani	47	85.1	61.7	**87.2**
Toms	33	**100.0**	90.9	**100.0**
Total	3907	88.2	72.9	**92.3**

Table 2. Sample sizes and accuracy rates. S: spectral information. T: temporal information. ST: spectral plus temporal information.

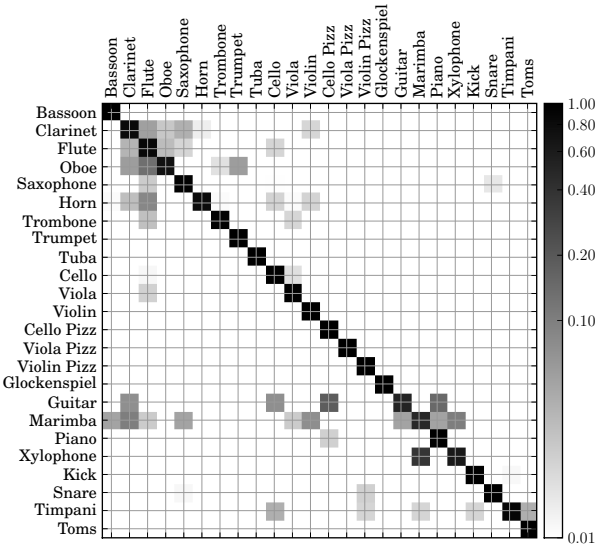

Figure 6. Classification accuracy using spectral information. Row labels: True class. Column labels: Estimated class. Average accuracy: 88.2%.

5.1 Spectral Information

As a control experiment, we evaluate the classification ability of spectral features using MFCCs. From each column of $\mathbf{A}$, we extract 32 MFCCs with center frequencies logarithmically spaced over 5.3 octaves between 110 Hz and 3951 Hz. From the 3907 32-dimensional feature vectors, we evaluate classification performance through ten-fold cross validation.

Fig. 6 illustrates the confusion matrix for this experiment, and Table 2 shows the accuracy rates for each class. The average of the 24 accuracy rates is 88.2%. We notice some understandable misclassifications. For example, 18.5% of guitar samples are misclassified as cello pizzicato and 14.8% are misclassified as piano. 5.5% of clarinet samples and 13.6% of oboe samples are misclassified as flute. 10.3% of marimba samples are misclassified as xylophone. In general, these spectral features can accurately classify the drums, brass, and string instruments. However, accuracy is poor among the woodwinds and pitched percussive instruments. Some of these misclassifications are due to an imbalance in the sample size of each class. Despite its ability to improve the average accuracy rate, the reduction of class imbalance in supervised classification is beyond the scope of this paper.

5.2 Temporal Information

Next, we evaluate the classification ability of temporal features using the proposed feature extraction algorithm with the parameters shown in Table 1. One feature vector $\mathbf{z}$ is computed for each temporal NMF atom as described in Section 4. Like the previous experiment, we evaluate classification performance through ten-fold cross validation among the 3907 32-dimensional feature vectors.

Table 2 shows the accuracy rates for each class. The average accuracy rate is 72.9%. Fig. 7 illustrates the confusion matrix for this experiment. We observe that temporal features alone do not classify instruments as well as spectral features. Nevertheless, for 11 out of the 24 classes, accuracy remains above 80%. In particular, there are very few misclassifications between percussion instruments and non-percussion instruments. Most misclassifications occur within instrument families, e.g., cello and viola, bassoon and clarinet, and guitar and piano.

5.3 Spectral Plus Temporal Information

Finally, we evaluate the classification performance when concatenating spectral and temporal features. The features extracted during the previous two experiments are concatenated to form 3907 64-dimensional feature vectors. Table 2 shows the accuracy rates, and Fig. 8 illustrates the confusion matrix. The total accuracy rate is **92.3%**. Temporal information improves classification accuracy for 16 of the 24 instrument classes along with the overall accuracy. Accuracy improves most for the string pizzicato, percussion, brass, and certain woodwind instruments. The remaining misclassifications occur mostly within families, e.g., clarinet and flute, and guitar and piano. For isolated sounds, this experiment verifies the hypothesis that temporal information can improve instrument recognition accuracy over methods that use only spectral information.

6. CONCLUSION

From the experiments, we conclude that a combination of spectral and temporal information can improve upon those instrument recognition systems that only use spectral information. The proposed method extracts temporal infor-

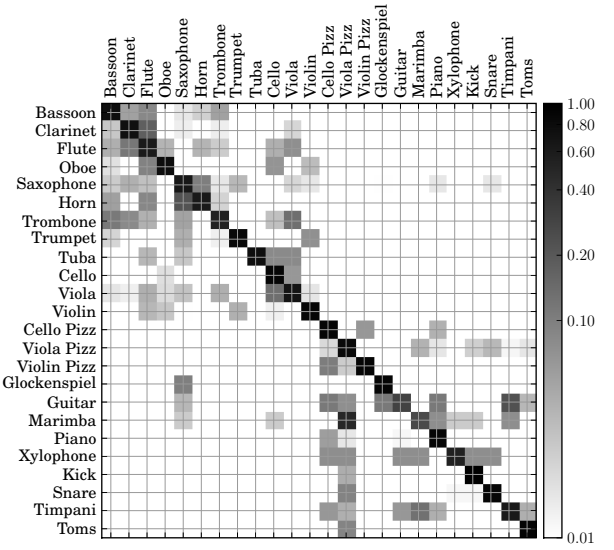

Figure 7. Classification accuracy using temporal information. Row labels: True class. Column labels: Estimated class. Average accuracy: 72.9%.

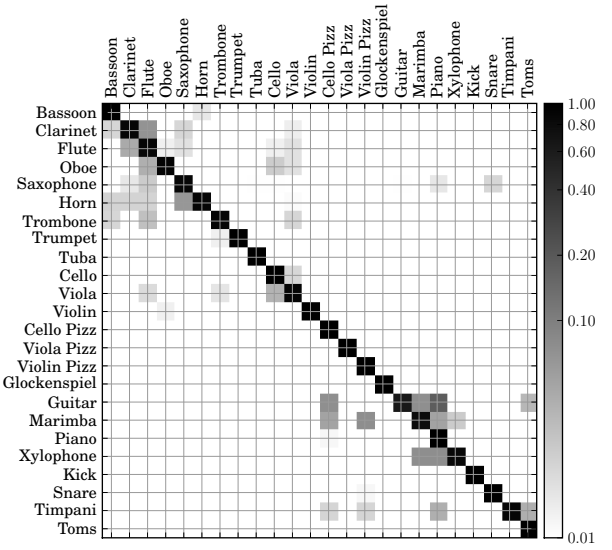

Figure 8. Classification accuracy using spectral plus temporal information. Row labels: True class. Column labels: Estimated class. Average accuracy: **92.3%**.

mation using a multiresolution gamma filterbank which parameterizes each temporal dictionary atom by its most prominent attack times and decay rates. Like the cortical representation, the spectral and temporal dictionary atoms generated by NMF provide a complete timbral representation of musical sounds. However, unlike the cortical representation, each of these dictionary atoms typically represent an individual musical note, thus facilitating music instrument recognition further.

We have already begun an investigation of the proposed method for both solo melodic excerpts and polyphonic mixtures. Also, because the proposed method classifies each individual NMF atom by instrument, we are investigating the use of the proposed method for source separation by grouping, emphasizing, or removing atoms that correspond to chosen instruments.

7. REFERENCES

[1] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001–. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[2] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoustical Soc. America*, vol. 118, no. 2, pp. 887–906, Aug. 2005.

[3] A. Eronen, "Automatic musical instrument recognition," Master's thesis, Tampere University of Technology, Oct. 2001.

[4] L. Fritts, "Musical Instrument Samples," Univ. Iowa Electronic Music Studios, 1997–. [Online]. Available: http://theremin.music.uiowa.edu/MIS.html

[5] F. Fuhrmann, M. Haro, and P. Herrera, "Scalability, generability, and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music," in *Proc. Intl. Soc. Music Information Retrieval Conf. (ISMIR)*, 2009, pp. 321–326.

[6] P. Herrera-Boyer, A. Klapuri, and M. Davy, *Signal Processing Methods for Music Transcription*. New York: Springer, 2006, ch. 6, pp. 163–200.

[7] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 2, pp. 424–434, Feb. 2008.

[8] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–. [Online]. Available: http://www.scipy.org

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Adv. Neural Information Processing Syst.*, vol. 13, Denver, 2001, pp. 556–562.

[10] R. Lyon and S. Shamma, "Auditory representations of timbre and pitch," in *Auditory Computation*, H. L. Hawkins, Ed. Springer, 1996, ch. 6, pp. 221–270.

[11] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 3, pp. 920–930, May 2006.

[12] "Freesound Project," Music Technology Group, Univ. Pompeu Fabra. [Online]. Available: http://www.freesound.org

[13] "Free Sound Samples – OLPC," One Laptop per Child. [Online]. Available: http://wiki.laptop.org/go/Sound_samples

[14] F. Opolko and J. Wapnick, "McGill University Master Samples," McGill Univ., 1987.

[15] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Appl. Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2003, pp. 177–180.

[16] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.