Instrument Recognition in Polyphonic Mixtures Using Spectral Envelopes

Jay Biernat
University of Rochester
jbiernat@ur.rochester.edu

ABSTRACT

Instrument recognition in polyphonic music is a difficult task in computer audition. Many current methods approach this problem by first attempting to separate the timbre features among the sources present in the mixture using source separation, multi-pitch estimation, or note-onset techniques. Instrument (timbre) recognition then proceeds on these separated features. This study proposes another method of instrument recognition which does not rely on first separating out the timbre features within a mixture. The method explored in this study uses spectral envelope templates to identify instrument timbre. This is done in a polyphonic mixture by finding the combination of spectral envelope templates that most closely approximates the spectral envelope in frames of the polyphonic audio. The method was tested using mixtures of two instruments playing sustained notes. However, our experiments did not achieve results significantly higher than the guess rate for instrument identification. We suspect that the poor results were partially due to our method's peak-detection algorithm, which performed well when applied to frequency spectrums of isolated instruments but had difficulty detecting relevant peaks in polyphonic mixtures.

1. INTRODUCTION

Musical instrument recognition is important for many tasks in the field of computer audition, such as music information retrieval, music transcription, and source separation. For instrument recognition in monophonic audio, various methods achieving good results already exist, of which [1] gives a good overview. Instrument recognition in polyphonic mixtures, however, is still an active research area.

The task of instrument recognition is mostly a problem of timbre recognition, a property of sound that is not well-defined in terms of a physical quality as other features like pitch (frequency) and loudness (sound pressure) are. However, previous work [2-3] has shown that a sound's spectral envelope is a robust feature for identifying an instrument's timbre. Several studies have proposed methods making use of this feature for instrument recognition in polyphonic audio. Burred *et al* [4] employs onset detection to group together spectral frequencies of the same note, the evolution of which are then tracked over time and matched with different instruments. Wang *et al* [5] first detects the fundamental frequencies present in a frame and then uses an SVM to classify a "characteristic timbre vector" calculated from the fundamental frequencies' harmonics. Heittola *et al* [6] uses a source-filter model to first separate the instruments in a polyphonic recording before using MFCCs in a GMM classifier to perform identification.

These methods all employ some form of source separation among the timbre features before instrument recognition takes place. The goal of this study is to explore a method of timbre recognition in polyphonic mixtures that does not require separation of timbre features among the different sources. To do this, we propose detecting instrument timbres on a frame-by-frame basis using 'spectral envelope templates' modeled using non-linear functions. We determine which combination of spectral envelope templates have the smallest error in predicting the peak amplitudes in the polyphonic frequency spectrum for each frame. The templates producing the smallest error are returned as matches for that frame. Once this is done over all frames, the templates with the highest number of matches over the entire audio sample are chosen as the instruments present in that sample.

The feasibility of this method is tested using a limited set of instruments and test data consisting of short, ideal audio samples that include only the sustained portion of played notes. We ignore the onset and the offset of the notes so that we can focus on providing matches for sustained note spectral envelopes, a simpler task than if the evolution of the spectral envelope during onset and offset was to be taken into account. Even with a limited instrument set and ideally constructed polyphonic audio samples, our method does not achieve success rates above the guess rate. It is not clear whether the idea behind the proposed method or its implementation is the reasons for these poor results.

The following paper is organized as follows: In section 2.1 the method used to create the instruments' spectral envelope templates is described. The two methods for using these templates to detect instrument timbre on a frame-by-frame basis is discussed in section 2.2. In section 3 we describe the experiment and the dataset used in the study. Section 4 is a review of the experimental results, and section 5 discusses areas for improvement and future work.

2. METHODS

2.1 Spectral Envelope Approximation

To create a spectral envelope template for a specific instrument, a non-linear function is fit to the spectral peaks found in all the steady-state note spectra across an instrument's range.

First, a steady-state note spectrum must be calculated for each note. A spectrogram of an instrument playing a single note in isolation is calculated using frame lengths of 46ms windowed with a hamming window and a hop size of 23ms. The spectrogram's phase information is discarded, and the linear magnitude from five consecutive frames of the sustained portion of the note are averaged together to smooth out frame-by-frame irregularities. Then, the amplitudes in the averaged spectrum are divided by the sum of the squared amplitudes to normalize the spectrum's power:

$$\widehat{x_i} = \frac{x_i}{\sum_{n=1}^{N} x_n^2} \tag{1}$$

Equation (1) shows this normalization, where $\widehat{x_i}$ is the normalized amplitude of the *i*-th frequency bin, and N is the total number of frequency bins. The normalized amplitudes are then converted from a linear to a decibel scale.

The normalized spectrum is then fed to a peak-detection algorithm that returns all peaks detected in the spectrum. The algorithm searches each frequency bin along the frequency spectrum looking for local maxima. When a local maxima is found, it is only accepted as a spectrum peak if two additional conditions are met:

$$\sum_{n=-2,-1,1,2} \frac{(x_i - x_{i-n})^2}{4} > \lambda \tag{2}$$

$$\sqrt{(i - i_{prev\_peak})^2 + (x_i - x_{i_{prev\_peak}})^2} < \alpha \tag{3}$$

Equation (2) specifies that the average of the squared differences between the amplitude of the local maxima and the amplitudes at the four surrounding frequency bins (two on each side) should be greater than λ. This ensures that there is a large enough change in amplitude between the local maximum and the surrounding frequency bins and causes small local maxima that don't correspond to true peaks in the frequency spectrum to be discarded. Equation (3) specifies that the Euclidean distance between the detected local maximum and the previous spectral peak not be greater than α. This prevents large jumps between peak amplitudes, which cause the algorithm to discard maxima that do not contribute to a smooth spectral envelope. In our experiments, we found setting λ = 500 and α = 50 provided good peak-detection in the frequency spectrums of isolated instruments

After spectrum peaks have been detected for all steady-state notes over a portion of the instrument's range, all the peaks are plotted together on a single plot, and a nonlinear function is fit to them using the GRG nonlinear solver in Microsoft Excel©. This fitted function will act as the spectral envelope template that will be used to detect that instrument's timbre in spectrums of polyphonic audio. Note that for many spectral envelope templates (such as the trumpet spectral template shown in Fig.1) piece-wise nonlinear functions are used to more accurately fit the spectrum's envelope. The frequency ranges where different curves are used in the piece-wise function are chosen by inspection.

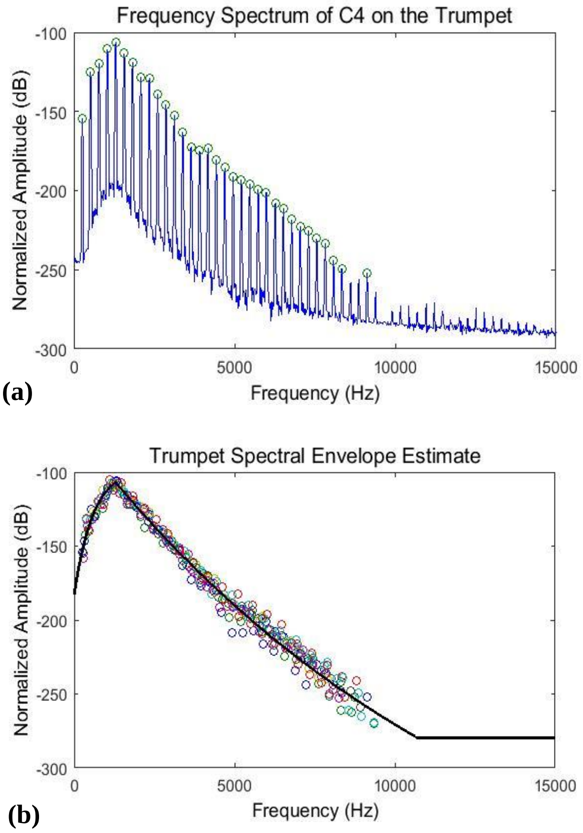

Figure 1. (a) The detected peaks in the frequency spectrum of a trumpet playing the note C4. **(b)** Detected peaks from all steady-state note spectra over one octave (C4-B4) of the trumpet's range. The dark line shows the function that has been that will be used as the spectral envelope template for the trumpet's timbre.

2.2 Frame-Level Timbre Detection

Two different algorithms for timbre detection are used in this study. Each algorithm is tested separately, but they are applied in the same manner. Each algorithm is applied frame by frame to a polyphonic audio sample, and for each frame, the algorithm returns the timbre matches found in that frame. The instruments with the highest number of total timbre matches over the length of the audio sample are labeled as the instruments present in the recording.

The first algorithm looks for an "overall fit" of the mixture's spectral envelope when compared with combinations of spectral envelope templates. The second algorithm checks for a "peak-by-peak" fit of each peak in the mixture's frequency spectrum with the amplitude of the corresponding frequency bin in each spectral envelope template. Both of these algorithms are described below.

In the following sections, the algorithms are explained assuming that there are only two instruments in an audio sample because this is a restriction we enforced in our test audio samples. However, both algorithms can be abstracted to work with audio containing larger numbers of instruments.

2.2.1 Overall Fit Method

For each frame of polyphonic audio, the frequency spectrum is calculated and normalized, converted to a decibel scale, and has its peaks detected as has been described in section 2.1. (Note: averaging over consecutive frames does not occur). After the peaks for the current frame have been detected, the overall fit algorithm finds timbre matches in the following way:

1. Create a 'possible fit' template from a combination of two spectral envelope templates
2. Determine the error between the possible fit template and the detected peaks of the polyphonic spectrum
3. Repeat steps 1 and 2 for all combinations of two spectral envelope templates
4. Determine which possible fit template has the smallest error from the detected peaks
5. Return the two spectral envelope templates whose combination created the possible fit template with the smallest error as "timbre matches" for that frame

This is repeated for each frame. The two instruments that have the highest total number of timbre matches over all frames are taken to be the two instruments present in that audio sample.

A 'possible fit' template is formed by taking the maximum of the two spectral envelope templates being combined

$$z_i = \max(x_i \, , y_i) \tag{4}$$

where z_i is the amplitude of the 'possible fit' template at frequency bin i, and x and y are the two envelope templates being combined. An example of a possible fit template using a combination of trumpet and saxophone envelope templates is shown in Fig. 2. Once the possible fit template is formed, the error between the possible fit template and the detected peaks in the current frame is calculated using the sum of the squared distances:

$$Error = \sum_{n=1}^{N} \left(a_{i_n} - z_{i_n}\right)^2 \tag{5}$$

where a_{i_n} refers to the amplitude at the i-th frequency bin where the n-th peak was detected, $z_{i_n}$is the amplitude of the possible fit template at the corresponding frequency bin, and N is the total number of peaks detected in the polyphonic spectrum.

The possible fit template with the smallest error determines the two timbre matches in that frame, and the two instruments with the highest number of matches over all frames are the instruments detected in that sample.

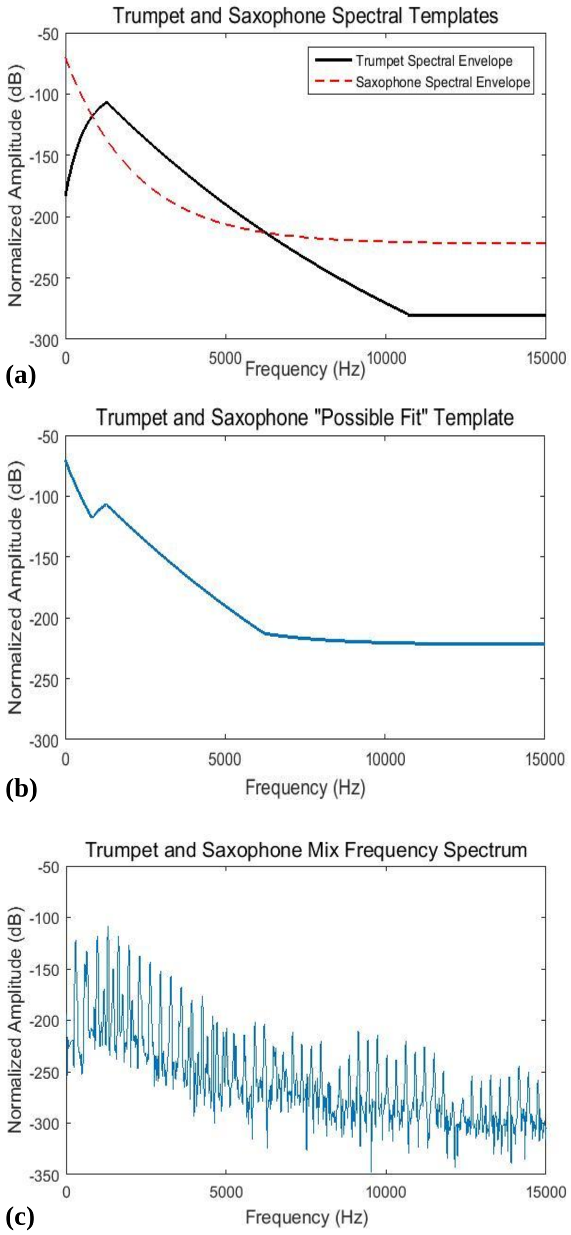

Figure 2. (a) The spectral envelope templates for the trumpet and saxophone. **(b)** The 'possible fit' template for the combination of trumpet and saxophone spectral envelopes. **(c)** The frequency spectrum of a frame of audio containing a trumpet and saxophone playing simultaneously.

2.2.2 Peak-by-Peak Method

The peak-by-peak method of matching timbre within frames finds the closest template match for each peak in the mixture spectrum. The algorithm iterates through each detected peak and finds the absolute difference between the amplitude of the detected peak and the amplitude at the corresponding frequency bin of each spectral envelope template. The algorithm then chooses the instrument template with the smallest distance as the best fit for that peak. The two instruments whose templates are matched to the most peaks in that frame are returned as the two instruments present in that frame.

Just as is done in the overall fit method, the two instruments that are returned the most throughout all frames in the audio sample are then identified as the two instruments in that sample.

3. EXPERIMENTS

Our method was tested using instrument audio samples from the University of Iowa's Music Instrument Samples (MIS)[1] database. Notes across an octave range of each of six different instruments were used both to create the spectral envelope templates and to create the test data. The six instruments used were: Bb clarinet, flute, French horn, oboe, saxophone, and trumpet. The range of notes used from each instrument, except for the flute, was the octave from C4 to B4. For the flute, the notes C5 to B5 were used. All note samples used were annotated in the database as being played with a *mezzo-forte* dynamic.

The spectral envelope templates for each instrument were created, as described in section 2.1, using the detected peaks from all notes across the chosen octave of the instrument's range. These templates were then used in both the overall fit and the peak-by-peak fit algorithms to detect which instruments were present in the test audio mixtures.

The tested audio samples were synthesized by combining note samples from different instruments into a single mono audio file. Each audio file is approximately one to two seconds in length and contains only two instruments each playing a single note. The onset and offsets of the notes are removed from the synthesized audio so that the mixture audio contains only the sustained portions of the notes. For simplicity, note onsets and offsets are not examined in this study because the spectral envelopes of note onsets and offsets are not consistent with the envelope templates obtained from the frequency spectra of sustained notes.

A total of twenty different test audio files were created. In fifteen of these, two different instruments are playing different pitches, and in five of these, the two instruments are playing the same pitch.

4. RESULTS

When evaluating the success of our method, both single matches and complete matches were kept track of. We defined a single match as one correctly identified instrument out of the two present in an audio sample. A complete match was defined as correct identification of both instruments present in an audio sample. The percentage of single and complete matches achieved for each algorithm are summarized in Table 1.

For both algorithms, the success rate achieved is no higher than simply guessing the correct instrument, which is 33% for a single match and 3% for a complete match.

	Overall Fit	Peak-by-Peak Fit
Single Match	40%	35%
Complete Match	5%	5%

Table 1. The percentage of successful single and complete matches achieved using the overall fit and peak-by-peak algorithms.

For both algorithms, there were two templates that were chosen disproportionately more than the rest. For the overall fit algorithm, this was the clarinet and saxophone templates, and for the peak-by-peak algorithm, the flute and oboe templates were favored.

5. CONCLUSIONS AND FUTURE WORK

The poor results achieved by our method may be due to a number of reasons. One part of our method that needs improvement is the peak-detection algorithm. Although the algorithm works well with peak detection in spectrums of isolated instruments, it has difficulty choosing the peaks most relevant to the spectral envelope in 'noisier' spectrums containing more than one instrument. Fig. 3 shows the peaks detected in a spectrum of a frame of audio containing a trumpet and saxophone.

Another factor that may be limiting the success of this method is the fact that the spectral templates are static. They currently cannot be adjusted to take into account the relative difference between the amplitudes of the two different instruments present in the audio. For example, although the audio samples of both the flute and the French horn are annotated as being played at *mezzo-forte*, test audio containing these two instruments has much more French horn presence. In the frequency spectrum of the mixture, the peaks of the flute's harmonics appear very small compared to the French horn's, which our algorithms do not take into account.

In order to fully examine the feasibility of using spectral envelopes to detect instruments in polyphonic mixtures, a better peak detection algorithm and a way to account for the relative loudness of instruments in mixtures should be explored in future work.

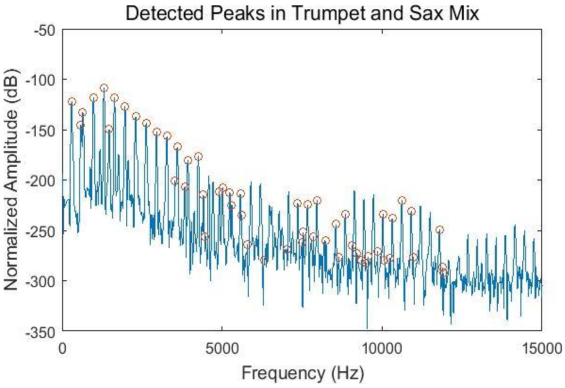

Fig 3. Detected peaks from an audio sample containing a trumpet playing E4 and a saxophone playing D4.

[1] http://theremin.music.uiowa.edu/MIS.html

6. REFERENCES

[1] P. Herrara-Boyer *et al*, "Automatic Classification of Pitched Musical Instruments," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York, NY: Springer (Science + Business Media LLC), 2006, pp.163-200.

[2] S. McAdams *et al*, "Discrimination of musical instrument sounds resynthesized with simplified spectrotemporal parameters," *The Journal of the Acoustical Society of America*, vol. 105, pp.882-897, Feb. 1999.

[3] J. J. Burred *et al*, "An accurate timbre model for musical instruments and its application to classification," in *Proc. 1st Int. Workshop LSAS*, Athens, Greece, 2006, pp. 22-32.

[4] J. J. Burred *et al*, "Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope," *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009, pp. 173-176.

[5] Y. Wang *et al*, "Automatic transcription for music with two timbres for monaural sound source," in *IEEE Int.Symp.Multimedia*, 2010, pp. 314-317.

[6] T. Heittola *et al*, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. 10th Int. Soc. Music Information Retrieval Conf.*, Kobe, Japan, 2009, pp. 327-332.