Instrument Recognition in Polyphonic Mixtures Using Spectral Envelopes

Jay Biernat

University of Rochester

INTRODUCTION

Instrument detection in monophonic recordings is a mostly solved problem in the field of computer audition [1]. Detection within polyphonic recordings, however, is much more difficult. When detecting timbre, most methods rely on some form of pre-processing, such as pitch estimation [2,3] or source separation [4], before timbre classification takes place. In this study, the goal is to explore a new method of timbre detection that does not rely on first accurately separating the different instruments present in the audio.

The spectral envelope of a sound is a robust feature for timbre discrimination and is relatively stable across an instrument's range [5]. When two instruments are present in the same frequency spectrum, the shape of the resulting spectral envelope is roughly the maximum of the two instrument's spectral envelopes. This study looks at the viability of using spectral envelope templates for detecting instruments in a polyphonic mixture.

Non-linear functions are used to model the spectral envelopes of instruments. These function templates are then used to compute the error between combinations of instrument templates and the detected spectral peaks in a polyphonic frequency spectrum. Unfortunately, this method does not produce results above the guess rate.

KEY IDEA

Use a nonlinear function as a template of an instrument's spectral envelope to detect that instrument's presence in a polyphonic sound recording. Detect the spectral peaks in a frame of polyphonic audio and compare the amplitude of each peak with the amplitude of an instrument's template at that frequency. The template combination that produces the smallest error when approximating the mixed frequency spectrum peaks will represent the instruments present in the polyphonic audio.

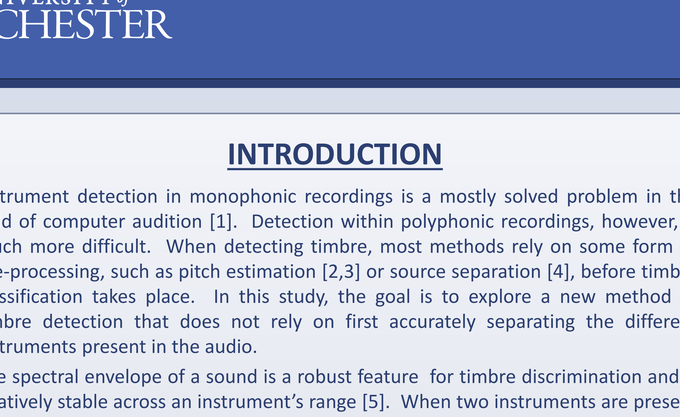

Fig. 1 The spectral envelope of a note played on a trumpet in isolation. The green data points show where the peak-detection algorithm has found significant peaks in the spectrum.

DATASET

The instrument note samples used to create the envelope templates and the test data were taken from the University of Iowa Music Instrument Samples (MIS)[1] database. One octave ranges from six instruments were used in the study: Bb clarinet, flute, French horn, oboe, saxophone, and trumpet. To create the test data, twenty short (1-2s) audio files were synthesized using the same note samples. Each audio file contains only two instruments, each playing one note, and does not include either the onset or the offset of the note.

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

METHOD

Envelope template generation

Peak Detection:

For a single, isolated note played by an instrument, an averaged frequency spectrum over the sustained portion of the note is calculated. The peaks in this frequency spectrum are then obtained using a developed peak-detection algorithm (**Fig 1**). This algorithm detects local maxima in the frequency spectrum and then determines whether the peak has a large enough change in amplitude to be considered significant. The algorithm also rejects peaks with a large difference in amplitude compared to the previous accepted peak to try to eliminate false peak detection.

Curve Fitting:

Non-linear, piecewise curves are fit to the envelope formed by the spectral peaks of a single instrument. This is done using the GRG Nonlinear solver in Excel (**Fig 2**).

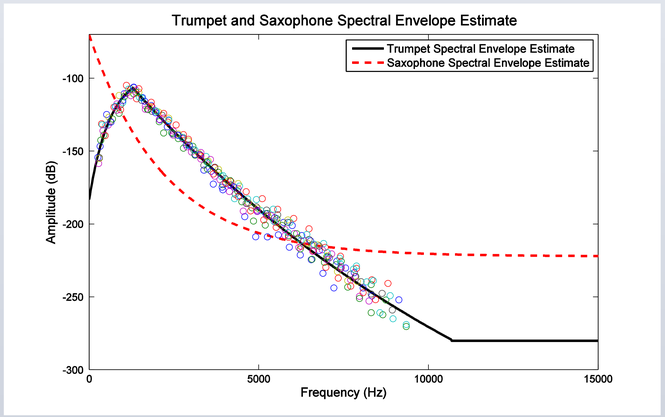

Fig. 2 All spectral peaks detected in trumpet notes C4-B4. The thick line is the fit (piece-wise) nonlinear curve that acts as the envelope template. The dashed red line shows the saxophone spectral envelope estimate for comparison.

Template matching

Two different methods of template matching are used. Each method is applied frame-by-frame to the test audio samples, and for each frame, the two most likely timbre matches are returned. The two instruments with the highest number of timbre matches over the length of the entire test sample are labeled as the instruments present in the recording.

Overall Fit:

Timbre envelope template functions are taken in pairs and combined to produce a "possible fit" template calculated by taking the maximum amplitude given by the two templates at each frequency bin. The squared error between the amplitude of each detected peak and the amplitude of the "possible fit" function at the same frequency bin are calculated and summed. The "possible fit" template with the lowest total error for that frame is returned, and the two instrument templates that are present in the possible fit template are returned as the instruments present in that frame.

Peak-by-peak Fit:

Each peak in a frame is examined separately. The error between a peak's actual amplitude and each of the template amplitudes at the same frequency is calculated, and the template with the lowest error for that peak is taken as a match. The two instruments with the highest number of peak matches for a frame are returned as the two instruments present in that frame.

RESULTS

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

A single match is one correctly identified instrument out of the two present in a sound file while a complete match refers to the correct identification of both instruments in a sound file.

Given that only six instruments were used to create test data set, and each audio sample has only two instruments, the peak-by-peak fit method performs no better than the guess rate for a single match (33%) or for a complete match (3%). The overall fit method does slightly better for single matches. The overall fit method seems to do a better job accurately identifying instruments with lots of upper harmonic energy, such as the trumpet and saxophone.

CONCLUSIONS

It is unclear as to whether the key idea or the implementation of the idea is the cause of the poor results obtained in the experiment. One major source of error is the poor performance of the peak detection algorithm when applied to polyphonic mixture (**Fig 3**).

While the algorithm works well with monophonic spectra, it has difficulty with peak detection in polyphonic spectra because of the often large amplitude jumps between peaks associated with different instruments and the overall more noisy-ness of the frequency spectrum.

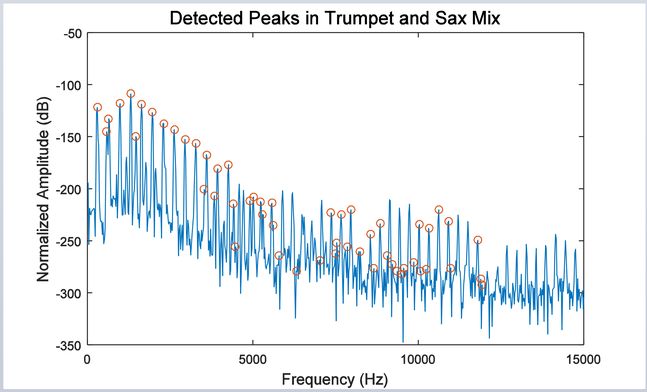

Fig. 3 Detected spectrum peaks from an audio file containing a trumpet playing E4 and a saxophone playing D4.

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] 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.

[3] T. Kitahara, M. Goto and H. Okuno, 'Musical instrument identification based on F0-dependent multivariate normal distribution', *2003 IEEE Int. Con. on Acoustics, Speech, and Signal Processing, 2003. Proc. of (ICASSP '03).*, vol. 5, pp. 421-4, 2003.

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

[5] 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.

ACKNOWLEDGEMENTS

Thanks to Dr. Duan for providing feedback and insight throughout the completion of this project; my classmates for their feedback; and the ECE Department for covering the cost of poster printing.