AUDIOVISUAL PARSING OF POLYPHONIC MUSIC

Madeline Cappelloni
University of Rochester Biomedical Engineering
madeline.cappelloni@rochester.edu

ABSTRACT

Hearing loss can have profound consequences on individuals, not only restricting their comprehension of speech, but also their enjoyment of music. Studies have shown that congruent visual cues can help to ameliorate the difficulties of comprehending a target auditory stream in complex environments. I will implement a visual system that provides temporally congruent visual stimuli to instruments in a polyphonic piece. I propose a simple non-negative matrix factorization (NMF) approach to source separation for extracting pitch and loudness features. The features will provide audiovisual coherence to the user interface. The user will be able to display a visual stimulus that represents the pitch and loudness of any given instrument in the piece they are listening to, helping them attend to that instrument. I achieve moderate performance of the source separation and propose further testing that can demonstrate the efficacy of this system for aiding people with hearing loss in music comprehension.

1. INTRODUCTION

Hearing impairments afflict individuals of a wide range of ages and can range from mild to profound hearing loss [1]. For these individuals, complex sound mixtures are particularly difficult to comprehend and there has been little effort to improve sensory aids that may help with polyphonic music comprehension. I propose a system that exploits the multisensory connections in the brain to improve polyphonic music comprehension with visual stimuli representing each source in the mixture.

Maddox et al showed that detection of a pitch modulation of a target sound in the presence of a masker was improved by a disk that changed its radius with the amplitude of the target sound [4]. Because the visual stimulus binds with the matching auditory stimulus, the brain is better able to perceive changes not only in the common feature, but also in orthogonal features. By providing a visual stimulus for each instrument in a musical piece, the listener can enhance their perception of any given instrument by attending to the matching visual stimulus.

By choosing features of the visual and auditory stimuli that match pre-existing associations, the binding across modalities can be strengthened. Brightness of a visual stimulus has be shown to be strongly associated with both auditory loudness and pitch [5]. Further associations between color and pitch exist in synaesthetes [2]. As such, the brightness and color of each shape will be dictated by the pitch of the instrument they represent. Loudness and size are implicitly associated by the propensity of distant objects being both small and quiet. The loudness of each instrument will therefore dictate the size of the corresponding shape.

Because the source separation is effectively computed by the brain, source separation is needed only to improve the accuracy of feature estimation. I will employ a simple non-negative matrix factorization [3] to separate instruments in a polyphonic piece, extract pitch and loudness of each instrument, and construct a visual interface to help individuals better enjoy polyphonic music.

2. METHODS

All analysis and presentation code was written in Python 2.7. The user interface requires the expyfun package in order to run (see github for code).

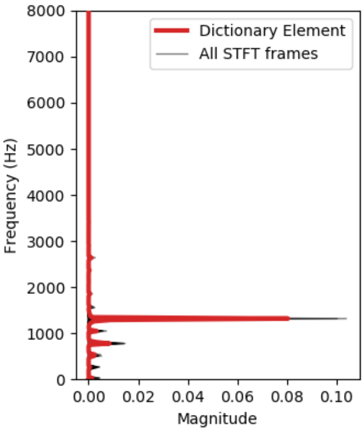

Figure 1. Demonstration of dictionary selection. The average of a subset of frames captures the frequency domain shape of non-silent frames.

2.1 Dataset

We will use recordings of notes from the Philharmonia Orchestra in order to form dictionaries for the each instrument in the polyphonic mix. In order to create a full set of instrument notes, we used only the audio file label forte normal for all available notes. For each note, I took the STFT for all frames with a Tukey window (α=0.25) and then selected the dictionary element for each frame by taking the average of frames that exceeded half of the average energy of the frame that had the most energy (Figure 1). By only taking

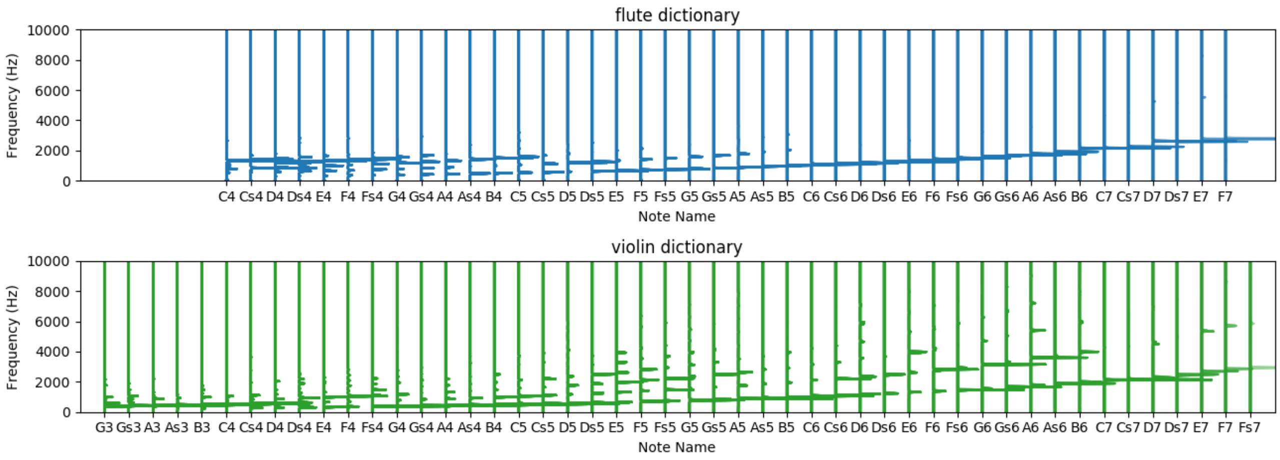

Figure 2. Dictionaries for flute (top) and violin (bottom).

a subset of frames, I did not include the frames which were silent. Importantly, because of the importance of audiovisual synchrony, the audio frame length was determined by the frame rate of the monitor (audio sampling rate divided by visual sampling rate). Complete dictionaries for two instruments are shown in Figure 2.

Audio files for analysis were taken from a free library of open source music, Musopen. Initial testing was performed on “Duet for Flute and Violin in G Major” by Franz Anton Hoffmeister.

2.2 Source Separation

In order to separate individual instruments, non-negative matrix factorization was performed using dictionaries constructed from notes played by each instrument. I made a magnitude spectrum of the piece of music using the STFT parameters as before. The algorithm used is included in python’s scikit learn package. H was initialized as the concatenated dictionaries of all instruments, and only W was updated with each iteration. Sparseness criteria (with a degree of 5) was implemented such that components from both instruments would not be activated for each note. The objective function for optimization is

$$Z = \sum_{i,j} (X - WH)^2_{ij}$$

where Z is minimized by a projected gradient solver. Optimization is completed when the objective function reaches a value of 10^{-6}.

2.3 Pitch Estimation

The pitch of each instrument is estimated as the pitch as the index of the dictionary at the maximum activation for the given instrument in the given frame. In order to refine this estimation, I multiplied the activations by a score informed instrument-specific cost function to the activations of each pitch. The cost function was a gamma function with α= 100 that was manually shifted to be centered around the median ground truth pitch. This should aid in pitch estimation because it is unlikely that the lowest and highest pitches in the range of the instrument will be played. The cost function decreases the amplitude of these pitches since they are more likely to be errors.

Because of the inherent noise in the estimate, I applied a forward backward 4th order lowpass filter with a cutoff frequency of 600 Hz. Indices were assigned based on the range of all instruments (i.e. half step intervals from the minimum pitch across instruments to the maximum pitch across instruments).

2.4 Loudness Estimation

Loudness was estimated as

$$L = \sqrt{\sum_i W^2_{ij}}$$

and normalized by

$$L_{norm} = L * \frac{0.01}{\sigma_L}$$

where σ_L is the standard deviation of the loudness. This was then transformed to a log scale for perceptual relevance. I filtered these estimates as above.

2.5 Visual Interface

Visual shapes were equally spaced polygons. The number of sides of the polygon was different for each instrument, with lower pitched instruments being assigned fewer sides than higher pitched instruments. Each polygon was randomly assigned a monochromatic color map ranging from a dark hue to white (see Matplotlib’s sequential colormaps). The different shapes and hues of the polygon emphasizes the visual differences between sounds.

The height of the polygons was determined by the absolute pitch of the recording (pitches of both instruments normalized to integer values between 0 and 99). The brightness of each polygon was determined by the relative pitch

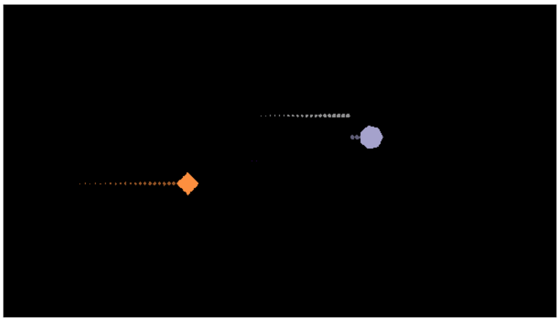

Figure 3. Visual interface with two simulated instruments playing.

of the instrument (each instrument's dynamic range normalized to integer values between 0 and 99). Size of the polygons is proportional to the loudness in dB. The size of the polygon is set to zero when it is below a threshold. The history of each instruments pitch and loudness is logged by identical shapes of decreased size and opacity that trail off to give the appearance that the canvas is traveling to the left. See Figure 3.

Visual representations of any instrument can be turned on or off independently by clicking near the shape. If the visual for a given instrument has been turned off, only a small white version of the shape with a 0.1 degree radius is shown to give an indication of where to click to turn it back on. The small "off" shape does not have dynamic shape, color, or position.

2.6 Testing

Testing data was constructed by taking the notes used to generate the dictionaries and stringing them together in a rudimentary melody. One melody was composed in C Major and a second was composed with random pitches. I then calculated the normalized RMS of each frame with the isolated audio for one instrument to act as a ground truth. The ground truth pitch was based on the labels of each audio clip.

3. RESULTS

The feature extraction was not particularly successful (see Figure 4) for the piece that was initialized with a key. However, the piece without a key (and therefore fewer harmonic relationships was) had limited accuracy as well (see Figure 5). In the case of the C Major piece, the pitch cost function is different for each instrument, reducing the errors that result in pitches overlapping. In addition to increasing pitch estimation accuracy, the pitches are more separated. This helps to improve visual separation of the two auditory streams.

In the case where there is no key, the cost function for pitch refinement hinders rather than helps the pitch estimation. Because the pitches of each instrument are chosen randomly and their ranges overlap entirely, the cost function for both instruments is the same and cannot aid in the separating the two instruments.

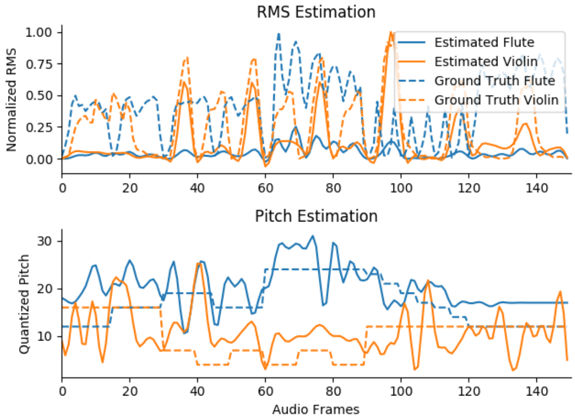

Figure 4. Results of the pitch and RMS extraction for C Major piece and cost function applied to pitch estimation.

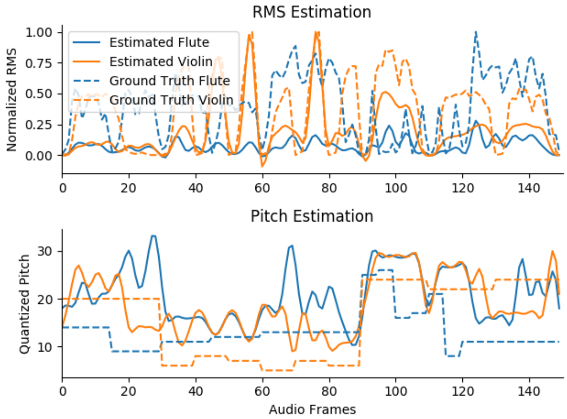

Figure 5. Results of the pitch and RMS extraction for piece without a key and cost function applied to pitch estimation.

Observations suggest that the performance of the visual system with either of these estimates are poor and not significantly affected by the application of the cost function. When using the ground truth RMS and pitch, the visual system performs well and gives the impression of being helpful in auditory comprehension.

See the supplemental video for a demonstration of the

Accuracy	RMS	Pitch
Key (with cost)	33%	28%
(without cost)	-	18%
No Key (with cost)	31%	13%
(without cost)	-	23%

Table 1. Accuracy of the RMS and pitch estimates from NMF relative to the ground truth. Tolerances of RMS and pitch accuracy were 0.1 and 3 respectively.

visual system using the ground truth pitch and RMS values.

4. DISCUSSION

Using NMF to separate sources necessitates that this be an offline program. Before separating the sources, prior knowledge of the instruments' identities is required. Further processing time is required to a converge on a solution. The system may be modified to allow for multitrack MIDI files to alievate the burden on any audio processing algorithms since these contain the required pitch and loudness information for each instrument.

Failure of feature extraction can be attributed to the basic nature of the NMF implementation. In order to correct this, further sparsity and temporal continuity constraints can be added. It is also possible that other source separation methods would yield more accurate results in estimating the RMS. Additionally, multipitch streaming methods may be employed. It should be noted that these techniques may be more computationally expensive and not provide profound benefits.

The testing is likely an overestimate of performance since the testing data was derived directly from the NMF dictionaries whereas real music would not. I observed a decrease in performance when using a real piece of music relative to the testing audio. This is solely a failing of the audio processing, as the observed performance of the visual interface is excellent when using the ground truth values. It should be noted that this evidence is anecdotal and needs to be formalized.

In order to truly assess the value of this system, it is important to test the benefits to human perception, rather than just the performance of the feature extraction. This could be achieved by asking normal hearing listeners to perform a melody matching task. In each trial, the listener would hear a clip of a polyphonic piece of music and be cued to attend to a particular instrument. Then, they would hear two melodies and be asked to identify which was played by the cued instrument. We would test three conditions: no visual system, the proposed system with all visuals on, and the proposed system with only the attended visual on. Validation of the proposed system would be achieved if a significant improvement was measured between task performance in the no visual and attended visual conditions. The all visual condition is may or may not offer a benefit and is therefore important to test in order to instruct users on proper use. While human testing was beyond of the scope of this work, it is necessary for evaluating the system.

In addition to proper testing, the system could benefit from more sophisticated audio processing techniques and visual imagery. For example, timbral features could be extracted and represented by the hue of the visual stimuli, or the shape of the visual stimuli could be mapped onto another auditory feature. Ultimately, the goal will be to maximize audiovisual coherence.

5. CONCLUSION

With growing insights into our reliance on multisensory information to perceive our environment, this project contributes to a body of work that takes technology into the multisensory realm. As a field, we should continue to exploit the power of multimodal information.

6. REFERENCES

[1] Y. I. Carroll, J. Eichwald, F. Scinicariello, H. J. Hoffman, S. Deitchman, M. S. Radke, C. L. Themann, and P. Breysse. Vital Signs: Noise-Induced Hearing Loss Among Adults - United States 2011-2012. *MMWR. Morbidity and mortality weekly report*, 66(5):139–144, February 2017.

[2] Kosuke Itoh, Honami Sakata, Ingrid L. Kwee, and Tsutomu Nakada. Musical pitch classes have rainbow hues in pitch class-color synesthesia. *Scientific Reports*, 7, December 2017.

[3] Chih-Jen Lin. Projected Gradient Methods for Nonnegative Matrix Factorization. *Neural Computation*, 19(10):2756–2779, October 2007.

[4] Ross K Maddox, Huriye Atilgan, Jennifer K Bizley, and Adrian KC Lee. Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *eLife*, 4.

[5] Lawrence E. Marks. On Associations of Light and Sound: The Mediation of Brightness, Pitch, and Loudness. *The American Journal of Psychology*, 87(1/2):173–188, 1974.