Audiovisual Parsing of Polyphonic Music

Madeline S. Cappelloni
Department of Biomedical Engineering, University of Rochester, Rochester, NY, USA.

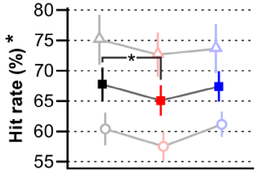
1. Introduction

BROADER IMPACTS

- Hearing loss is common among adults, particularly with aging
- Such individuals struggle to comprehend sounds in complex auditory environments and can have impaired ability to enjoy music
- This work will offer a visual system to aid individuals in listening to polyphonic music

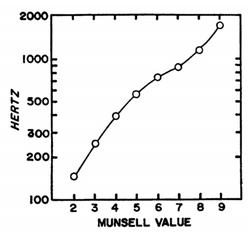
PSYCHOPHYSICAL MOTIVATION

- Temporal coherence of a visual stimulus to one auditory stream in a mix enhances attention to the matching stream [1]



- **Visual matches target stream**
- **Visual matches distracter stream**
- **Visual matches neither stream**

- This effect should be bigger for more salient audiovisual connections
- When participants were asked to match a tone with a grey shape, they matched low pitch tones with darker shapes and high pitch tones with brighter shapes [2]
- Brightness was also found to be associated with loudness of the tone
- Experience with objects at varying distances produces an association of loudness and size



SOURCE SEPARATION

- Non-negative matrix factorization (NMF) has been successfully used to separate audio sources in a mix
- Separated audio is generally not clean, but is useful in extracting features from sources
- We use methods from scikit learn[3]

This work uses supplemental visual stimuli such that the user performs "source separation" on polyphonic music rather than relying on a computational method.

References

1. 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.
2. 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.
3. Chih-Jen Lin. Projected Gradient Methods for NonnegativeMatrix Factorization. Neural Computation, 19(10):2756–2779, October 2007.

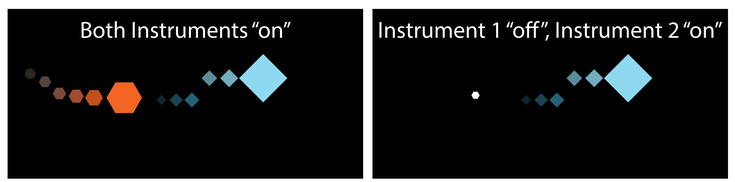
2. Methods

AUDIO PROCESSING

- Dictionaries for each instrument were generated from recordings of 1 second duration notes over the full range of the instrument
- I performed NMF on the magnitude spectrogram of the piece - each column of the spectrogram represents an audio frame that is the length of one frame presentation on the monitor (e.g. for 30 fps and 44100 Hz audio sampling, frame length is 1470 samples)
- Scikit learn's NMF with projected gradient descent with sparseness criteria was used, only updating the activation matrix
- All entries of each instrument's activation matrix were used to estimate loudness
- I found the largest activation for each time frame and deem this the pitch in that frame
- To improve pitch estimate, I applied a score informed cost function (gamma function centered near the mean pitch) to the activations
- RMS and pitch estimates were smoothed with a 4th order 600 Hz low pass filter

VISUAL INTERFACE

- Visual entries are regular polygons
- A different shape and hue are assigned to each instrument to maximize their independence from each other
- Sizes of the shapes are updated with the normalized RMS
- Brightnesses and heights of the shapes are updated with the estimated pitch
- "History" of the instrument trails off to the left of the main shape with decreasing size and opacity

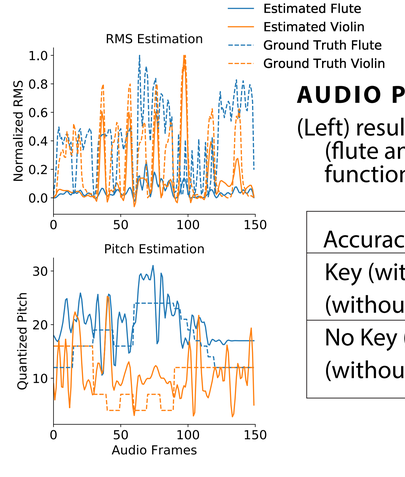


- User may turn each shape "off" or "on" at will -- "off" shapes are just a smaller white version of the shape with no history and no dynamic properties

TESTING

- Constructed test materials from recordings of notes that were used to make the dictionaries
- Each note was a random length with a cosine window applied to the end for a smooth transition
- Some test pieces were initialized such that they had a key, others were constructed with random notes that were not harmonically related
- Ground truth loudness was estimated from the isolated streams and ground truth pitch was estimated as the labels of each of the recorded notes

3. Results



AUDIO PROCESSING

(Left) results of testing with the test piece (flute and violin) with key and cost function

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

VISUAL INTERFACE

- Anecdotal evidence suggests improved auditory comprehension using ground truth RMS and pitch but not when using estimated RMS and pitch

4. Discussion

AUDIO PROCESSING

- Techniques for source separation leave much to be desired
- May call for more sophisticated separation and pitch estimation algorithms
- Potential for online approach to be developed

VISUAL INTERFACE

- Additional auditory features could be respresented in visual shapes (e.g. timbre)

TESTING

- In order to accurately assess the performance of the system, human music comprehension has to be tested with and without the visual system
- I propose a future experiment: engage normal hearing listeners in a melody identification task in background noise. Listeners will perform the task with 2 randomly interleaved conditions: no visuals, and proposed system. We can compare performance in the conditions to see if the proposed system offers a significant benefit.

Multimodal processing is both important in human perception and a critical concern for the advancement of computer audition.