ZHIYAO DUAN RESEARCH STATEMENT

The objective of my research is to design computational systems that are capable of analyzing and processing auditory scenes. I call this **Computer Audition**. Although humans can naturally analyze auditory scenes into meaningful events, and consciously or passively respond to them, computer audition remains a challenging problem and once it succeeds, it will represent a fundamental advance in **Audio Signal Processing** and **Intelligent Systems**. Successful computer audition systems will also improve our ability to access and manipulate audio data, enabling new multimedia applications and interactions for experts and novices alike.

Auditory scenes are difficult to parse because source signals composing them overlap and interfere with each other at the same time and frequency. For example, when one is walking on a street, the auditory signal that he/she receives at any time and frequency may be a mixture of the rustle of leaves, people chatting, and a car passing by. Furthermore, the ways that these sources mix together and the source signals themselves are constantly changing. Imagine recognizing objects in a visual scene where every object is half-transparent (sound sources overlap), objects change their appearance (sound sources change over time) and even disappear or reappear unexpectedly (sound sources become silent and active).

A successful computer audition system must be able to access source signals in spite of the other interfering sources. This challenge raises two fundamental questions: 1) How do we discover and separate a sound event from an audio mixture? 2) What sound events belong to one source versus other sources? To answer these questions, I performed research at three levels: 1) sound source tracking (e.g. identifying footsteps and streaming them together); 2) audio source separation (e.g. extracting vocals from music, removing noise to enhance speech); and 3) leveraging external information (e.g. a musical score or speech transcript) to improve separation and manipulation of sound objects and sources.

In the following three sections I will outline my work in these three areas. This work has been funded by NSF Information & Intelligent Systems (IIS) grants 0643752 and 0812314, and has been published in collaboration with researchers at Microsoft Research, Adobe Systems Inc. and Gracenote Inc.

SOUND SOURCE TRACKING

Identifying sound objects and streaming them into sources is fundamental in analyzing an auditory scene. For example, in analyzing an auditory scene of walking up stairs, we hear multiple individual sounds separated by silence. We segment these sounds and identify them as footsteps. We also realize that these sounds go together, i.e. they are from the same person. Empowering computers with this ability is very difficult because sound objects are “hidden” by other interfering sources (e.g. footsteps + humming), and streaming them together requires finding out their invariant characteristics, i.e. timbre, which is physically not well defined.

I have done significant work on sound source tracking by focusing on multi-pitch tracking, i.e. estimating a pitch trajectory (stream) for each harmonic source (e.g. harmonic musical instruments, vocals and speech) in an audio mixture. Here pitches correspond to sound objects. Accurate, reliable multi-pitch tracking would be of great utility in many tasks, including music transcription, speech recognition, melody-based music search, etc. While most existing methods identify pitches, few of them attempt streaming them according to sources. In addition, all existing methods are designed to work for either music or speech, but not both or mixtures containing both. A general method that can deal with different types of harmonic sources would significantly broaden its applications.

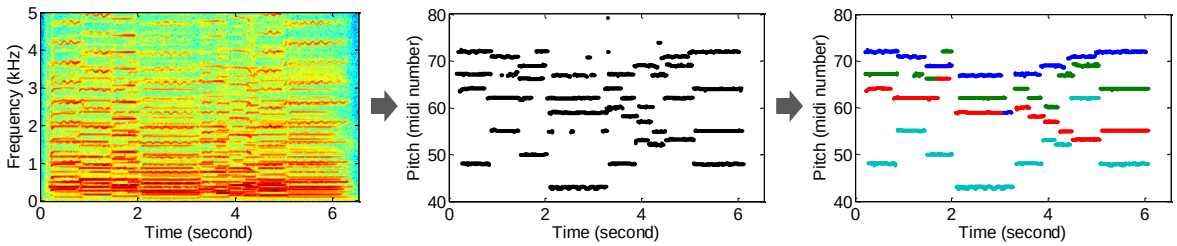

Figure 1. Multi-pitch tracking first identifies concurrent pitches at each time frame of the audio spectrogram, then streams pitches across frames according to sources. Pitch trajectories of different sources are shown in different colors.

I have developed a new probabilistic method to identify concurrent pitches and their number (polyphony) in each time frame, suitable for different types of harmonic sources [1]. Example output is shown in the center panel of Figure 1**.** Model parameters are trained for different categories of harmonic sounds (e.g. music, speech, bird singings) instead of specific sources. This method has shown success on both music and speech data, better than existing methods that were specifically designed for each sound category.

I then developed the first method to stream pitch estimates across time into pitch tracks according to sources [2]. Example output is shown in the right panel of Figure 1. This method can deal with both music and speech, and does not require pre-training source models. I frame this problem as constrained clustering of pitch estimates, where each cluster corresponds to a source. The clustering objective is to minimize the timbral inconsistency within each cluster. To solve this problem, I developed a novel constrained clustering algorithm, since existing algorithms could not cope with the problem's unique properties [3].

A key issue in streaming pitches (and sound objects in general) of the same source is finding their invariant timbral features, directly from the mixture. Existing features for speech (e.g. MFCC, LPC) can only be calculated from single-talker speech, and requires source separation beforehand for multi-talker scenarios. **I developed a novel timbral feature, which can be calculated for each talker directly from a multi-talker speech mixture [4].** This feature may open an avenue for many tasks in multi-talker speech processing such as speech/speaker recognition, prosody analysis, etc.

AUDIO SOURCE SEPARATION

One step forward from source tracking is source separation, i.e. recovering source signals from the mixture. It is of great utility for many further processing tasks, including talker identification, speech recognition, post production of existing recordings, and structured audio coding. I focus on separating multiple sources from a single-channel mixture. In this case, no peripheral information such as multi-channel spatial cues can be used, leaving the core of auditory scene analysis to be addressed.

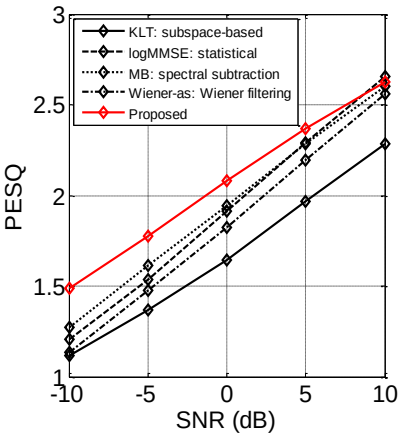

Figure 2. The proposed method outperforms four categories of classical speech enhancement algorithms in non-stationary noise environments, with different input SNR conditions. PESQ is a broadly used objective measure for speech quality.

When prior analysis of the individual sound sources composing the mixture is not possible, the problem is considered blind, or *unsupervised*. **I developed an unsupervised method to separate musical instruments and vocals in music [5]**, based on harmonic structure (relative amplitudes of harmonics) analysis. I found that harmonic structure is approximately invariant for sounds of each instrument within a narrow pitch range, but is significantly different for different instruments. All possible harmonic structures in the mixture were extracted and then clustered into several clusters, each of which corresponds to

an instrument. An average harmonic structure model was calculated for each cluster and was used to separate its signal. This method outperforms a Nonnegative Matrix Factorization-based method on a music dataset.

In some scenarios, when prior analysis is available for some sources (e.g. background noise in teleconferencing) but not for the others (e.g. speech of attendee), the problem is considered *semi-supervised*. Semi-supervised separation based on probabilistic latent component analysis (PLCA) has shown success in many scenarios, especially for non-stationary inharmonic sources. However, existing algorithms only work offline. They cannot be applied in online/real-time applications such as speech enhancement (separating speech from noise) in teleconferencing. **I developed an online algorithm for PLCA-based semi-supervised separation [6]**. Experiments on speech enhancement show that it outperforms four categories of classical algorithms in non-stationary noise environments [7], see Figure 2. Where dealing with non-stationary noise is still an obstacle in speech enhancement, the proposed algorithm might lead to a breakthrough.

LEVERAGING EXTERNAL INFORMATION FOR SOURCE MANIPULATION

Humans often leverage external information to analyze auditory scenes. For example, reading a musical score helps musicians enjoy and study a symphony. Leveraging textual, visual and other kinds of information allows a computer audition system to accomplish tasks that could not be accomplished by processing audio alone.

I focus on leveraging musical score information to improve music audio source separation and manipulation. To do so, one needs an alignment (mapping) between score objects (e.g. notes, chords) and audio signals. This is not trivial since the temporal dynamics of audio performances (instantiation) are often significantly different from those of the score (abstraction). In addition, rich timbral patterns in polyphonic audio blur its mapping to the abstracted score. **I developed a novel online approach to map a multi-instrument polyphonic music audio with its score [8].** This approach models an audio performance as a path in a two-dimensional performance space (state space). Finding the mapping is achieved by estimating the path (hidden states).

For semi-improvised music like Jazz, audio-score alignment is more challenging. The score, being more abstract, only specifies a basic melody, several chords and a musical form. The performer "composes" the details on the fly. **I developed the first approach that aligns polyphonic Jazz audio with their scores [9], without need for a separate microphone for each instrument.** Improvisations in these performances include note/chord/rhythm changes, and unexpected structural changes such as jumps and repetitions.

Based on audio-score alignment, **I further built an online score-informed source separation system, called *Soundprism* [10].** Given aligned score notes, the system first performs a more accurate pitch estimation of the audio sources. It then separates sources by constructing harmonic masks of their pitches. All the processing is performed in an online fashion. I envision a smart phone app implementation of it as shown in Figure 3. An offline version of the system has been implemented into an interactive audio editor. Users can select and edit the corresponding audio signal of a note or a source by clicking on the note in the score, and modifying it.

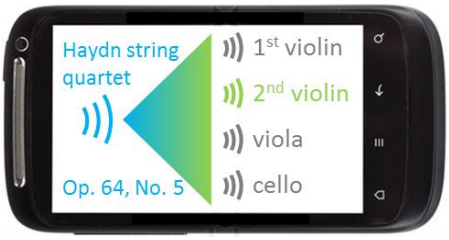

Figure 3. Soundprism can be implemented as a real-time smartphone app, which allows users to switch between enjoying the full performance and focusing on the "hidden" 2nd violin in a live concert.

RESEARCH AGENDA

I aim to build a computer audition system that is able to analyze complex auditory scenes composed by multiple concurrent sound sources. In the future, I will actively apply for research grants from the NSF and other funding agencies. I will also seek collaborations with other audio industries such as Dolby Home Theater, Apple Siri and Google Music. This section outlines some future opportunities that I am excited to pursue.

Tracking general sound sources. The proposed multi-pitch tracking system has shown success in identifying and streaming pitches of harmonic sources. A lot of environmental sounds like footsteps, however, are inharmonic and do not have a pitch. The ability of tracking general sources would allow the computer audition system to work in more natural scenarios. Would the "identifying-streaming" strategy still work? If so, what are the invariant features to stream the sound objects? One possible direction is to investigate the temporal dynamics of sound objects. Footsteps of the same person have similar temporal patterns due to his/her typical gait. Also, two footsteps following too close or too apart would not belong to the same person.

Source separation without pre-training source models. In many scenarios, isolated recordings for pre-training source models are not available for all (or part) of the underlying sources in a mixture. Source separation in these scenarios requires exploring inherent self-organizational cues (timbral, temporal and semantic) of the sources. This is a fundamental problem in source separation that I would like to address. Besides this, another route is to pre-train an alternative source model from a similar but available source (e.g. a violin solo), then adapt it to separate the underlying source in the mixture (e.g. a viola). The difficulty, however, is to find an appropriate alternative source model automatically, without manual recognition of the underlying sources.

Integrating other media with audio. Soundprism integrates musical score with audio. In the future, I will try to integrate other media especially visual information with audio. This will help analyze complex auditory scenes which are difficult if only use audio information. For example, mouth localization has been shown helpful for speech recognition in noisy environments. Gesture recognition of performers will help recognize instrumental sources in a symphony, and guide cameras to automatically focus on the corresponding performers.

REFERENCES

[1] **Zhiyao Duan**, Bryan Pardo and Changshui Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio Speech Language Process.*, vol. 18, no. 8, pp. 2121-2133, 2010.

[2] **Zhiyao Duan**, Jinyu Han and Bryan Pardo, "Harmonically informed multi-pitch tracking," in Proc. *International Society on Music Information Retrieval conference (ISMIR)*, 2009, pp. 333-338.

[3] **Zhiyao Duan**, Jinyu Han and Bryan Pardo, "Song-level multi-pitch tracking by heavily constrained clustering," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 57-60.

[4] **Zhiyao Duan**, Jinyu Han and Bryan Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE Trans. Audio Speech and Language Process.*, under review.

[5] **Zhiyao Duan**, Yungang Zhang, Changshui Zhang and Zhenwei Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Trans. Audio Speech Language Process.*, vo. 16, no. 4, pp. 766-778, 2008.

[6] **Zhiyao Duan**, Gautham J. Mysore and Paris Smaragdis, "Online PLCA for real-time semi-supervised source separation," in Proc. *International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), Lecture Notes in Computer Science 7191*, pp. 34-41, 2012.

[7] **Zhiyao Duan**, Gautham J. Mysore and Paris Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments," in Proc. *Interspeech*, 2012.

[8] **Zhiyao Duan** and Bryan Pardo, "A state space model for online polyphonic audio-score alignment," in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, 197-200.

[9] **Zhiyao Duan** and Bryan Pardo, "Aligning improvised music audio with its lead sheet," in Proc. *International Society for Music Information Retrieval Conference (ISMIR)*, 2011, pp. 513-518.

[10] **Zhiyao Duan** and Bryan Pardo, "Soundprism: an online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Process.*, vol. 5, no. 6, pp. 1205-1215, 2011.