ZHIYAO DUAN RESEARCH STATEMENT

The objective of my research is to design computational systems that are capable of understanding auditory scenes (i.e., Computer Audition) and to enable new kinds of interaction with audio. This interdisciplinary research not only draws from but also advances multiple research fields including Audio Signal Processing, Machine Learning, Human-Computer Interaction, and Music Cognition.

Together with our collaborators, my group has proposed novel research directions and made pioneering contributions. These directions include **query by example of audio databases** (e.g., searching for sounds using vocal imitation as the search key), **audio-visual scene understanding** (e.g., extracting a person's voice from a cocktail party recording by leveraging the person's lip movement information), and **human-computer collaborative music making** (i.e., designing computational systems that can improvise music together with humans through natural audio-visual interaction). Our contributions along these directions range from algorithm design (audio representation learning, audio-visual signal modeling and search algorithm design) to system integration (development of sound search engines and interactive music systems).

The reminder of this document provides highlights of our contributions along these three novel directions. Additional projects and contributions that are not covered here include **speech enhancement** [1-3] (noise reduction and bandwidth extension), **speech emotion analysis** [4-7] and **sound event detection** [8, 9]. See www.ece.rochester.edu/projects/air/projects.html for a full list of projects with examples and demos.

Audio-Visual Scene Understanding

Humans use the concert of all the five senses to understand the world; human audition rarely functions in isolation from other senses especially vision. Seeing lip movements of a talker helps us understand what is being said in a noisy environment. Hearing crowd cheering helps us understand the status of a basketball game. The computer audition and computer vision communities, however, were segregated due to various reasons. Take the music information retrieval (MIR) community as an example, although gestures and facial expressions of musicians have been shown to be even more important than sound for musical expressions and audience engagement, prior research had not paid much attention to the visual modality. Recognizing this as a great opportunity, **I initiated a collaboration with image/vision researchers on audio-visual scene understanding.** Our audio-visual work spans from music performance analysis to speech processing, and to general audio-visual scenes.

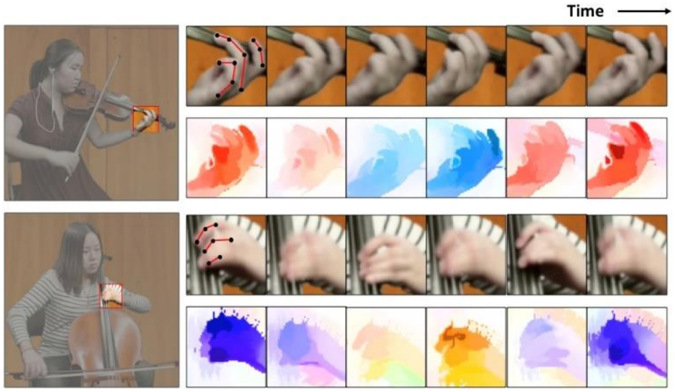

Figure 1. Hand motion detection with optical flow technique. Color maps show direction (color) and velocity (intensity) of

On MIR research, my lab did groundbreaking work on audio-visual music performance analysis. **We created the first audio-visual multi-instrument multi-track music performance dataset** that proves valuable for multiple existing and new research tasks [10].

We proposed a new research task named "audio-visual source association": associating sound sources or score tracks with musicians' body movements in music performance videos. This association is critical for visually informed source separation, and would enable applications such as isolating a musician's part by clicking on the musician in YouTube videos. Our initial work for string ensembles won **the best paper award** in the 2017 Sound and Music Computing (SMC) Conference [11]. Our follow-up work extends this association modeling to all kinds of string, woodwind and brass instruments in Western chamber music, without the need of knowing the kind of instruments beforehand [12]. It models audio-visual correspondences between note onsets and body/finger motion, and between pitch fluctuation and vibrato motion of musicians in music ensembles.

On performance technique analysis, **we proposed the first video-based method for vibrato analysis for polyphonic string music** [13]. Vibrato is a common playing technique characterized by a periodic pitch fluctuation. Existing methods analyze vibrato rate and depth by analyzing the pitch contour. In polyphonic performances (e.g., string quartet), however, pitch estimation itself is a challenging problem. For string instruments, vibrato is produced by the left hand rolling motion which changes the length and tension of the vibrating string. Our method uses the optical flow technique to estimate this hand rolling motion (see Figure), detects vibrato notes, and further analyzes vibrato rate and depth. Experiments show that the vibrato note detection accuracy is significantly improved from the traditional audio-based method (see Figure 2). This work won a **best paper nomination** in the 2017 International Society for Music Information Retrieval (ISMIR) Conference.

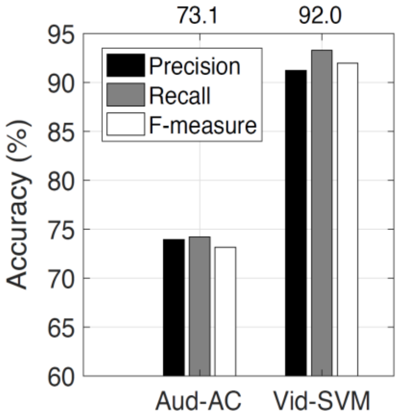

Figure 2. Vibrato detection accuracy comparing audio-based and the proposed video-based method for polyphonic string music.

The cocktail party problem, i.e., separating sound sources from an audio mixture is a fundamental problem in computer audition. **We proposed a novel audio-visual deep clustering (AVDC) method to leverage both audio and visual cues to separate the signals of multiple speakers [14].** This AVDC model is trained on audio-visual speech mixtures and their isolated sound sources to learn audio-visual time-frequency (T-F) embeddings that capture audio-visual association. During separation, this embedding is calculated for each speaker at each T-F bin, and T-F bin clustering is then performed to achieve separation. This method is shown to outperform two state-of-the-art audio-based and three audio-visual separation methods. Further analyses show that the AVDC model generalizes across different numbers of speakers between training and testing and shows some robustness when visual information is partially missing.

Our audio-visual work also includes the **generation of talking faces from input speech**, which would help speech comprehension for hearing impaired population and in noisy environments. We first proposed a system to generate identity-neutral 2D/3D talking face landmarks from input speech in real time [15, 16]. We further proposed a system to generate photo-realistic talking face videos from these generated face landmarks for an arbitrary person given the person's single face image [17, 18]. We are now working on an **end-to-end system** that bypasses the face landmark generation stage to directly generate photo-realistic talking face videos for an arbitrary person from speech spoken by other people. This system is robust to noise in the speech signal, and is able to generate emotion expressions when the speech is emotional. We foresee that this disruptive technology may pose challenges on privacy and security, and we plan to do research on its counterpart: automatic detection of generated fake videos.

Our audio-visual work is supported by the National Science Foundation grant No. 1741472, "BIGDATA: F: Audio-Visual Scene Understanding." The talking face generation work was also supported by a UofR AR/VR pilot award.

Human-Computer Collaborative Music Making

We are in a world where the interaction between humans and machines is becoming deeper and broader, but this interaction can hardly be called collaboration: The role of machine is assistive at best, and is not nearly equal to that of humans. Developing systems that allow us to truly collaborate with our increasingly important partners – machines, is one of the main missions in the research of cyber-human systems. In this project, we aim to advance human-computer collaboration by designing systems that allow humans and machines to play music together. State-of-the-art research along this line is on automatic music accompaniment, where the system follows human performer's tempo and renders synthesized or pre-recorded music accompaniment. This interaction is far from the way that human musicians collaborate with each other. We propose to design **systems that allow humans to collaborate with machines in the way that human musicians collaborate among themselves.**

Great collaboration relies on deep mutual understanding. To achieve our goal, machines need to understand human musicians' performance. Machines also need to know how to "perform" instead of mechanically playing back pre-recorded tracks to allow humans better understand them. Furthermore, machines need to understand musical knowledge and compositional skills to improvise music with humans. These three aspects – **perception, performance, and theory** – form the core skills of musicianship. Our lab has made significant contributions on all of the three aspects (see below).

On music perception, we made contributions on music transcription (converting music audio into musical score), score following (aligning music performance with musical score in real time) and audio-visual music performance analysis. For highlights, we developed **a highly accurate piano transcription algorithm for the context-dependent setting** using convolutional sparse coding [19, 20]. This algorithm, once trained on the specific piano and acoustic environment to be transcribed, achieves significantly higher transcription accuracy than the state of the art, albeit the presence of background noise and strong room reverberation. A US patent (9779706) is issued for this work. Furthermore, as existing music transcription systems only convert music audio into the MIDI representation, we proposed **a system to transcribe polyphonic music audio into MIDI and then into human-readable music notation** [21], and we designed **the first metric for evaluating music notation transcription** [22].

We also developed **a score following technique for piano that is robust to the audio-score mismatch due to the sustained effect**: the sound of a note lasts longer than what is notated in the score due to legato articulation, sustain pedal usage and room reverberation [23]. This approach increasingly improves the alignment accuracy and robustness over state-of-the-art baselines as the sustained effect becomes stronger. Finally, as described in the previous section, **we pioneered the research on audio-visual music performance analysis**, including source association [12], vibrato analysis [13], and multi-pitch estimation and tracking [24].

On music performance, we designed **the first system that renders a skeleton pianist playing the input music in real time (**Figure 3**)** [25]. The skeleton makes expressive body movements (e.g., leaning forward before playing heavy chords) that are often indistinguishable from those of the human pianist who played the music. While sound synthesis has been investigated for decades, to our best knowledge, **this was the first work on visual music performance rendering**. Further investigation along this direction would enable many novel applications in music education, entertainment, and AR/VR.

On music theory modeling, **we developed a music language model for multi-part music**, which models the relation between a position of the music score and its context [26]. This model is part-invariant, i.e., it can process/generate music scores consisting of an arbitrary number of parts using a single trained model. This model is used in melody/bassline harmonization and music generation. In a baseline harmonization experiment, we compared the machine harmonization with student harmonization at the Eastman School of Music, through blind, objective (grading) and subjective (listening) tests. Results show that our algorithm achieved comparable ratings to students on two of the three basslines [26]. See Figure 4 for objective evaluation results.

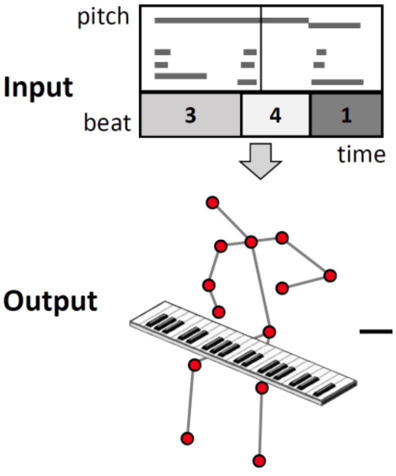

Figure 3. The pianist skeleton rendering system.

To our best knowledge, this was the first systematic comparison between machine composition and human composition of music in the literature.

Our ongoing work on music theory modeling and music generation is on **real-time interactive improvisation**. We have developed a human-computer interactive duet improvisation demo system that responds to human playing with a counterpoint performance in real time. The current algorithm behind this system is based on maximum likelihood prediction, and we are improving it by formulating the problem as reinforcement learning, which is commonly used in learning agent-environment interaction such as autonomous driving.

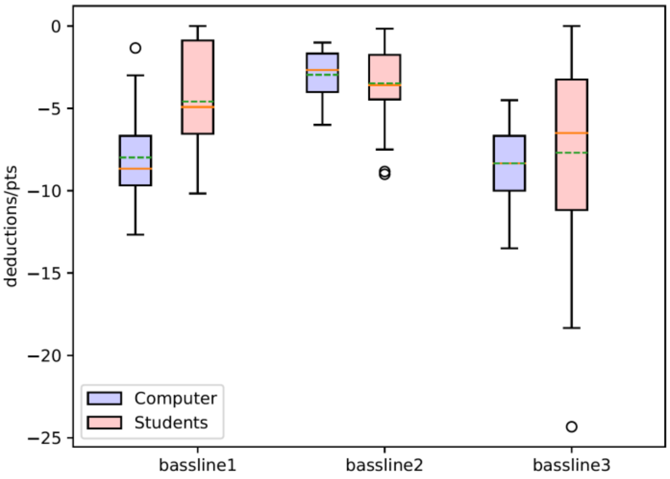

Figure 4. Objective evaluation (grading) on bassline harmonization by our algorithm and by music students. Higher values are better.

Moving forward, we will integrate the three parts of work – perception, performance, theory/composition – into the design of a human-computer collaborative music making system with audio-visual interaction. We also plan to design a physical embodiment of such system, leading to human-robot collaborative music performances. This work is supported by the National Science Foundation grant No. 1846184, "CAREER: Human-Computer Collaborative Music Making."

Query by Example of Audio Databases

Text-based search engines have profoundly improved the way people access information online. For audio databases, however, text-based search is often problematic. Many audio files have metadata that does not describe the audio content (e.g., "30 minutes of street scene, Chicago Aug 24, 2012"). Files labeled with content-relevant tags often have non-specific tags that make them indistinguishable from many other files (e.g., "Barking dogs" returns 1134 files on www.freesound.org). Even when a specific file is known to contain the desired content, it is typically not indexed with tags throughout the file, forcing the user to listen through the file to find the desired segment.

To address these problems, **we proposed a novel idea for audio search – using audio examples (vocal imitation in particular) as the query key** (Figure 5): A user vocalizes the audio concept in mind and the system retrieves audio recordings that are similar, in some way, to the vocal imitation. This approach complements the traditional text-based search, and would make search easier, faster, and possible in cases where text tags are not available.

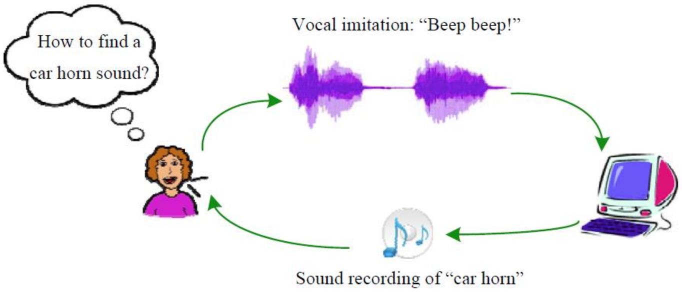
How to find a car horn sound?
Vocal imitation: "Beep beep!"
Sound recording of "car horn"

Figure 5. Sound search by vocal imitation.

There are two main challenges in designing vocal-imitation-based sound retrieval systems: 1) What feature representations should be used for the vocal imitation queries and database audio files, such that acoustic characteristics that humans use to imitate are emphasized while surface-level differences are downplayed? 2) What matching algorithms would work with the feature representations to discern target sounds from irrelevant sounds, given a query?

To answer these questions, we explored different techniques and system designs, from using engineered features such as mel-frequency cepstral coefficients (MFCC) to automatically learning features from data with stacked auto-encoders (SAE), from using pre-defined similarity measures such as cosine similarity and Kullback-Leibler (K-L) divergence to automatically learning these measures using neural networks [27]. Recently, we proposed a **novel Siamese-style neural network architecture called TL-IMINET** [28], as shown in Figure . It uses two convolutional neural networks (CNN) to extract feature representations from the log-mel spectrograms of the imitation query and a sound candidate in the database, and then uses a fully connected (FC) network with three layers to predict the similarity. TL-IMINET is trained on positive (relevant) and negative (irrelevant) pairs of vocal imitations and sound recordings to learn what features to extract and how to predict the similarity.

As many neural-network-based methods, TL-IMINET is data hungry, however, large amounts of imitation-sound pairs are expensive to obtain. To address this problem, we adopted the **transfer learning** idea to pre-train CNN feature extraction networks on relevant tasks (spoken language classification and environmental sound classification) where large amounts of data are available. Experiments show that TL-IMINET outperforms our previous state-of-the-art system IMISOUND on the average ranking of the target sound, thanks to the new network architecture and transfer learning [29].

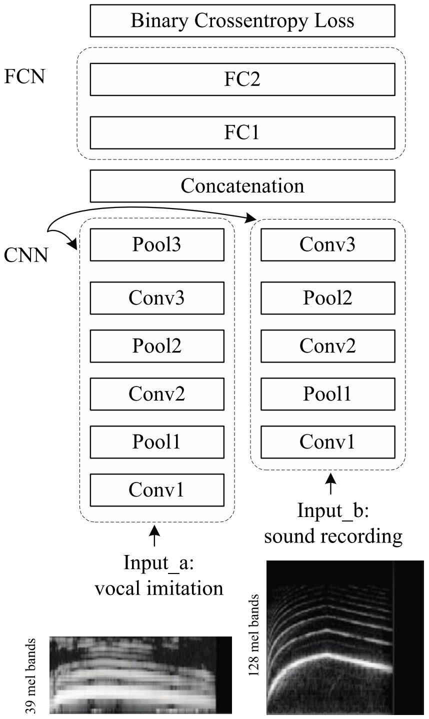

Figure 6. The proposed TL-IMINET neural network architecture. The two CNN networks extract feature representations and the FCN network predicts similarity.

Interpretation of deep neural networks remains an important open problem in machine learning. As the network architecture becomes more complex, it is more difficult to understand how the network functions. To better understand the feature representations and similarity measures learned by TL-IMINET, we used activation maximization to **visualize and sonify** input spectrograms that maximize the activation of certain neurons in each CNN layer and FCN layer. We found that the **CNN networks pay attention to delicate patterns** (e.g., bird chirping and water flowing like sounds) from the input vocal imitation and sound recording spectrograms, and the similarity measure learned by the **FCN network emphasizes the similar temporal evolution** and downplays the timber differences between the vocal imitation and sound recording.

We further developed a vocal-imitation-based sound search engine named *Vroom!* using the TL-IMINET algorithm and created a copyright-free database with ~5000 sound effects for this search engine. This will be **the first public sound search engine** taking vocal imitation queries. We also conducted a small-scale user study which showed that vocal-imitation-based search demonstrated significant advantages over text-based search for certain categories of sounds (e.g., synthesized sounds) in terms of the search efficiency and ease-of-use ratings [30]. **We are now conducting a large-scale user study** on Amazon Mechanical Turk to further assess the effectiveness of this novel way to sound search over the traditional text-based approach. Moving forward, we plan to integrate these two search approaches and scale to large databases with long audio recordings.

This work is funded by the National Science Foundation grant No. 1617107, "Algorithms for Query by Example of Audio Databases."

Selected References

[1] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman, "Front-end speech enhancement for commercial speaker verification systems," *Speech Communication,* vol. 99, pp. 101-113, 2018.

[2] J. Zhou, S. Chen, and Z. Duan, "Rotational reset strategy for online semi-supervised NMF-based speech enhancement for long recordings," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015, pp. 1-5.

[3] S. E. Eskimez, K. Koishida, and Z. Duan, "Adversarial Training for Speech Super-Resolution," *IEEE Journal of Selected Topics in Signal Processing,* vol. 13, pp. 347-358, 2019.

[4] N. Yang, J. Yuan, Y. Zhou, I. Demirkol, Z. Duan, W. Heinzelman*, et al.*, "Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification," *International Journal of Speech Technology,* vol. 20, pp. 27-41, 2017.

[5] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5099-5103.

[6] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. Heinzelman, "Emotion classification: how does an automated system compare to naive human coders?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2274-2278.

[7] S. E. Eskimez, M. Sturge-Apple, Z. Duan, and W. B. Heinzelman, "WISE: Web-based Interactive Speech Emotion Classification," in *Proc. SAAIP @ IJCAI*, 2016, pp. 2-7.

[8] R. Lu, Z. Duan, and C. Zhang, "Metric learning based data augmentation for environmental sound classification," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 1-5.

[9] R. Lu, Z. Duan, and C. Zhang, "Multi-scale recurrent neural network for sound event detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 131-135.

[10] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia,* vol. 21, pp. 522-535, 2018.

[11] B. Li, C. Xu, and Z. Duan, "Audio-visual source association for string ensembles through multi-modal vibrato analysis," *in Proc. Sound and Music Computing Conference (SMC),* pp. 159-166, 2017.

[12] B. Li, C. Xu, G. Sharma, and Z. Duan, "Online audio-visual source association for chamber music performances," *Transactions of International Society for Music Information Retrieval,* vol. 2, pp. 29-42, 2019.

[13] B. Li, K. Dinesh, G. Sharma, and Z. Duan, "Video-based vibrato detection and analysis for polyphonic string music," *in Proc. International Society for Music Information Retrieval Conference (ISMIR),* pp. 123-130, 2017.

[14] R. Lu, Z. Duan, and C. Zhang, "Audio–Visual Deep Clustering for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 27, pp. 1697-1712, 2019.

[15] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Generating Talking Face Landmarks from Speech," in *Proc. International Conference on Latent Variable Analysis and Signal Separation*, 2018, pp. 372-381.

[16] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Noise-resilient training method for face landmark generation from speech," *IEEE/ACM Trans. Audio, Speech & Language Processing,* Under Review.

[17] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," *in Proc. European Conference on Computer Vision (ECCV),* pp. 520-535, 2018.

[18] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7832-7841.

[19] A. Cogliati, B. Wohlberg, and Z. Duan, "Context-dependent piano music transcription with convolutional sparse coding," *IEEE/ACM Transactions on Audio, Speech and Language Processing,* vol. 24, pp. 2218-2230, 2016.

[20] A. Cogliati, Z. Duan, and B. Wohlberg, "Piano transcription with convolutional sparse lateral inhibition," *IEEE Signal Processing Letters,* vol. 24, pp. 392-396, 2017.

[21] A. Cogliati, D. Temperley, and Z. Duan, "Transcribing human piano performances into music notation," *In Proc. International Society for Music Information Retrieval Conference (ISMIR),* pp. 758-764, 2016.

[22] A. Cogliati and Z. Duan, "A metric for music notation transcription accuracy," *Proc. of International Society for Music Information Retrieval (ISMIR),* pp. 407-413, 2017.

[23] B. Li and Z. Duan, "An Approach to Score Following for Piano Performances With the Sustained Effect," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 24, pp. 2425-2438, 2016.

[24] K. Dinesh, B. Li, X. Liu, Z. Duan, and G. Sharma, "Visually informed multi-pitch analysis of string ensembles," *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 3021-3025, 2017.

[25] B. Li, A. Maezawa, and Z. Duan, "Skeleton plays piano: online generation of pianist body movements from MIDI performance," *in Proc. International Society for Music Information Retrieval Conference (ISMIR),* pp. 218-224, 2018.

[26] Y. Yan, E. Lustig, J. Vaderstel, and Z. Duan, "Part-invariant model for music generation and harmonization," *in Proc. International Society for Music Information Retrieval Conference (ISMIR),* pp. 204-210, 2018.

[27] Y. Zhang and Z. Duan, "Supervised and unsupervised sound retrieval by vocal imitation," *Journal of Audio Engineering Society,* vol. 64, pp. 533-543, 2016.

[28] Y. Zhang, B. Pardo, and Z. Duan, "Siamese Style Convolutional Neural Networks for Sound Search by Vocal Imitation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 27, pp. 429-441, 2019.

[29] Y. Zhang and Z. Duan, "IMINET: convolutional semi-siamese networks for sound search by vocal imitation," *in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* pp. 304-308, 2017.

[30] Y. Zhang, Y. Zhang, and Z. Duan, "Sound search by text description or vocal imitation?," *arXiv:1907.08661,* pp. 1-5, 2019.