ZHIYAO DUAN RESEARCH STATEMENT

My research focuses on **computer audition** (i.e., designing artificial intelligence (AI) algorithms to understand and generate various kinds of sounds) and its connections to computer vision, natural language processing, and human-computer interaction. I develop and apply machine learning and signal processing methods to solve fundamental problems in audio processing and multimodal learning, and I also collaborate with domain experts on applications in security, education, and healthcare.

I direct the **Audio Information Research (AIR) Lab** at the University of Rochester (UofR) since 2013. I have been supervising 15 UofR PhD students and 2 visiting PhD students. I took a sabbatical leave from June 2020 to January 2022 at Kwai Inc. to lead a research team in Seattle, working on multimodal interaction and creation. However, I continued supervising student research and growing my lab at UofR during my leave. Up to date, as PI or Co-PI, I have secured over **$6M** funding (**$3.7M** my share) from federal, state, industry and university funding sources. As senior personnel, I also contributed to other projects funded by NSF and NIH.

Since my promotion to Associate Professor with Tenure in 2020, my lab has been primarily working along four research directions: 1) **human-computer collaborative music making** (e.g., developing AI agents to improvise music with humans), 2) **toward an ecosystem of AI-powered music production** (e.g., bridging audio AI research and music production practices), 3) **speech security and privacy with generative AI** (e.g., speech anti-spoofing and anonymization), and 4) **personalized and adaptive spatial audio rendering** (e.g., predicting head-related transfer functions from ear and head geometry). The remainder of this document highlights our contributions along these directions. Additional projects with examples and demos can be found at www.ece.rochester.edu/projects/air/projects.html.

Human-AI Collaborative Music Making

We are in a world where the interaction between humans and AI is becoming deeper and broader. Developing intelligent systems that can collaborate with humans is one of the main missions of artificial intelligence and cyber-human systems research. In this project, we aim to advance human-AI collaboration by designing algorithms and systems that can play music together with humans. This project was funded by the National Science Foundation grant "CAREER: Human-Computer Collaborative Music Making" (1846184, 06/2019-05/2024, $499,219, sole PI) and gifts from ByteDance, Kwai, and Adobe.

Great collaboration relies on deep mutual understanding. To achieve our goal, AI needs to understand human musicians' performance. AI needs to acquire musical theory and compositional skills to compose or improvise. AI also needs to know how to "perform" instead of mechanically playing back pre-recorded tracks. These three aspects – **perception, theory, and performance** – form the core skills of musicianship. Our lab has made significant contributions along all these aspects since 2020.

On music perception, we made significant contributions to automatic music transcription (i.e., converting music audio into music notation). On piano transcription, we developed **an event-based transcription algorithm that directly outputs notes (including pitch, onset time, offset time, and velocity), skipping the frame-level pitch estimation** [1, 2]. This was an innovative approach as existing methods all adopted a two-stage approach: 1) estimating pitches within short frames (~50ms) of audio, and 2) connecting the pitch estimates into notes. The two-stage approach is limited because errors produced in the first stage are difficult to fix and the second stage loses rich context. Our proposed approach is elegant, addresses these limitations, and achieves the state-of-the-art results on piano transcription at the note level. This work received a **best paper nomination** at ISMIR 2024. Besides piano transcription, we also developed algorithms for guitar tablature transcription [3, 4], choral music transcription [5], and multi-instrument general multi-pitch estimation [6]. In particular, we proposed **the first method that leverages self-supervised learning (SSL) toward multi-pitch estimation** [6]. The key idea was to use SSL to learn invariances (to timbre and noise) and equivariances (to pitch shifting and time stretching) of the pitch concept from massive amounts of unlabeled music audio.

On music perception, we also made significant contributions to real-time rhythm analysis [7-9]. In particular, building on a sequence of prior efforts, we proposed BeatNet [10], **a state-of-the-art real-time**

joint beat and downbeat tracking algorithm for music. The key idea was to use a convolutional recurrent neural network (CRNN) to estimate beat and downbeat saliences in each audio frame, and then use particle filtering to decode the final beat and downbeat predictions in real time. Up to date, the code repository has received **75K downloads and 354 stars** on GitHub. Last year, we further improved this model to BeatNet+ [11] through auxiliary training and adaptation strategies. With this improvement, BeatNet+ is able to track beats and downbeats from percussion-less music audio and even isolated singing voices.

On music theory and composition, we developed **the first framework for modeling and generating multi-part music in Western modern staff notation** [12], while existing methods typically uses MIDI notation, which is not a natural notation format for human musicians. As such, our developed framework paved a path toward human-AI collaborative music composition. Furthermore, our framework models each music part as a sequence of measures and uses a hierarchical autoencoder architecture to decompose a measure into voices, chords, pitches and ties. Compared to the commonly used music representation as a sequence of notes, our framework respects the natural hierarchy of musical objects and significantly reduces the sequence length, making it easier to model long-term temporal dependencies. For online music generation, we proposed **the first algorithms for human-AI duet counterpoint improvisation**, using maximum likelihood estimation and reinforcement [13, 14]. We also employed inverse reinforcement learning to fuse the Western counterpoint style with the Chinese folk music style and to **generate countermelodies for Chinese folk melodies** [15].

On music performance, we are working on drummer animation driven by a drum set MIDI sequence in real time. While the work is not yet published, we have made good progress on drummer body movement and planning. On music synthesis, collaborating with Adobe, we developed a fast high-fidelity stereo vocoder to convert monophonic, low-resolution (e.g., 16 kHz) mel-scale magnitude spectrograms into stereophonic, high-resolution (e.g., 44.1 kHz) waveforms [16].

Building on the research in the abovementioned three core machine musicianship areas, we have developed a few human-AI collaborative music making systems. Here I highlight three projects on collaborative music making. **BachDuet** (www.bachduet.com) [13] (Figure 1) is a neural-network-based AI agent that can improvise duet counterpoint with a human musician in real time. The motivation of this project was to develop a counterpoint improvisation partner to help revitalize the improvisation culture in classical music education. A main design goal was to achieve a relatively equal role between the human musician and the AI agent in the collaboration. User studies confirmed we met this goal. Listening tests also suggested that the human-AI improvisations were not distinguishable from human-human improvisations. **Draw&Listen** [17] is another AI agent designed for people without musical training to compose music. It exemplifies how music AI can lower the barrier of music creation. It allows users to fill missing measures by simply drawing a pitch curve and additional note density curves. The agent converts these curves into notes that match the musical context. We chose this design because everyone, with or without musical training, understands concepts like “high” and “low” of pitch and “dense” and “sparse” of rhythm. **GrooveMate** is a drummer AI agent that can accompany a human musician (e.g., singer, jazz combo) in real time. This system will integrate the real-time rhythm analysis algorithm (BeatNet+ [11]), a drum track generation algorithm, and the drummer animation algorithm mentioned above.

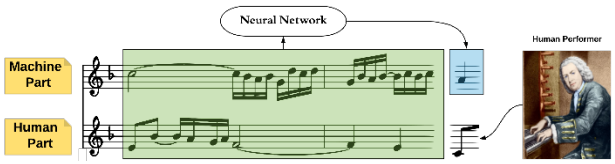

Figure 1. Illustration of BachDuet – a real-time AI agent for human-computer duet counterpoint improvisation.

For future research, I aim to develop embodied music AI agents through collaboration with robotics researchers. I plan to use specialized devices (e.g., player pianos) or build novel musician robots (e.g., drummer robots) and deploy our music perception, composition and performance algorithms to achieve human-robot collaborative music making. Key challenges to address include: 1) how to accommodate physical constraints of robots (e.g., delays, computational power) in the design of algorithms, 2) how to improve the interaction between human musicians and robotic musicians, and 3) how to integrate perception, reasoning and action into a unified learning framework for the robots to learn to collaborate with human musicians through rehearsals.

Toward an Ecosystem of AI-Powered Music Production

The impact of AI on music and audio may fundamentally change the music production industry. In 2019, I started to collaborate with researchers from education, music, philosophy, and computer science to think about broader issues of music AI research. Through an NSF planning grant (2026439, 09/2020-03/2022, $149,674, Co-PI), we proposed an exciting research direction: building an AI-powered music production ecosystem. The key idea is to **develop an open-source software framework to bridge music AI research and music production workflows, enabling musicians and audio AI researchers to collaborate**. This project is funded by the National Science Foundation grants (2222129 and 2222369, $1,800,000, 10/2022-09/2026, Co-PI taking share of $611,678) "Collaborative Research: FW-HTF-R: Toward an Ecosystem of Artificial Intelligence-Powered Music Production (TEAMuP)".

Nowadays, music and audio AI models are being developed on a weekly basis worldwide and they show fantastic performance on a large variety of old and new tasks. However, almost all these models only stayed in academia and were never used by musicians. On the other side, musicians primarily use digital audio workstations (DAWs), such as Logic, Pro Tools, REAPER, Cubase and Ableton Live, in their workflows. These tools, being commercial software, have very limited AI functionalities and are updated very slowly. AI researchers could deploy some of their AI models to these DAWs by writing plugins, however, most researchers do not have the skills or interest. Musicians also rarely have sufficient technical skills to try out AI models even if open-sourced. Our goal of this project is to remove this barrier by helping audio AI researchers easily deploy their models to DAWs ready for users to use.

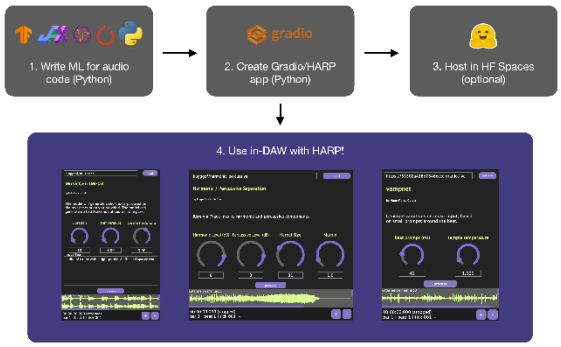

Figure 2. HARP lets DAW users access large state-of-the-art audio AI models with GPU compute from the cloud without breaking their within-DAW workflow.

In the past two and a half years, with failures of several approaches (e.g., VST protocol, Audio Random Access (ARA) protocol), we eventually converged to a solution that the open-source software framework is implemented as a stand-alone application but can be initiated from DAWs. This stand-alone application is named **HARP, standing for hosted, asynchronous, remote processing** [18, 19] (Figure 2). AI researchers can easily deploy their models to HARP by writing Python code. DAW users can send their audio data to HARP through the "bounce" function of their DAWs; after processing the audio data in HARP with AI models, they will see the audio data automatically transferred back to DAWs ready for other processing. Depending on the model size, computational complexity, and the need of real-time or offline processing, AI models can be hosted on remote servers like Hugging Face spaces or on the users' local machines. Up to date, multiple audio AI models have been deployed to HARP, including HPSS (source separation) [20], VampNet (music audio generation) [21], and Timbre-Trap (music transcription) [22]. In addition to audio AI models, HARP is also being developed to accept other types of input and output data formats including MIDI and text. MIDI models such as Music Transformer (MIDI-based music generation) [23] and melody harmonization [12] are being deployed, and audio-text models such as HTS-AT (audio captioning) [24] are also being deployed.

My lab has also been developing a web framework for users to play with interactive music systems, i.e., systems that can play music with users in real time. Like AI models, most interactive music systems only stayed in lab environments and never interacted with real users because of technical barriers on implementing production-level software. Extending from our BachDuet system [13], we developed **Euterpe, an open-source framework for deploying and using interactive music systems on the web**. Euterpe handles essential peripheral components of interactive music systems such as scheduler, user input interface, visual rendering, and audio rendering, freeing researchers to focus on their core algorithm development. Thanks to the web framework, Euterpe is naturally cross-platform. It supports both audio and MIDI interactions, grid-based (16-th note grid) and event-based music representations, and both call-and-response and simultaneous-play interaction paradigms. With Euterpe, deploying interactive music systems becomes much easier. For example, the PianoGenie system [25] which used 1000 lines of JavaScript code to develop now only needs less than 200 lines of code.

Speech Security and Privacy with Generative AI

The past few years witnessed significant advances of generative AI (e.g., ChatGPT, DALL-E, Sora). Generative AI for speech (i.e., speech synthesis) poses great challenges and opportunities for security and privacy. On the one hand, speech synthesis techniques such as voice cloning challenge existing speaker verification systems. Since 2020, my lab has been working on speech anti-spoofing and audio deepfake detection with the support from the New York State Center of Excellence (CoE) in Data Science grants ($239,655, 07/2021-06/2025, sole PI) and gift from Meta ($50,000 + Audiobox licence, 08/2024-07/2025, sole PI). These projects aim to develop and deploy fundamental technologies for speech anti-spoofing with our industry partner, IngenID, a Rochester-based company on voice biometrics. On the other hand, voice conversion and morphing techniques can help people protect their voice privacy. My lab has been working on controllable voice conversion, expressive text-to-speech (TTS) synthesis and voice morphing. Our research is funded through the IARPA Anonymous Real-Time Speech (ARTS) program (UofR share of $1.25M, 09/2024-09/2027, sole PI) collaborating with researchers from Honeywell, UT Dallas, and Texas A&M. I have also been collaborating with researchers from neuroscience, psychiatry and nursing on applying speech processing techniques to healthcare problems.

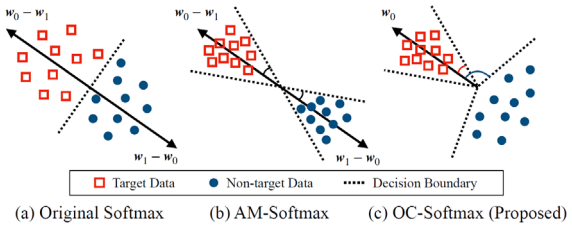

Figure 3. Illustration of the proposed one-class learning idea (c) compared with existing binary classification ideas (a, b).

Synthetic speech detection aims to classify synthetic speech utterances from bona fide utterances. It becomes increasingly important in speech security. The key challenge is to generalize to unseen attacks, i.e., utterances synthesized with techniques that are not used to train the detection system. Previous methods viewed the problem as binary classification treating both classes equally, but it intrinsically was not able to generalize to unseen attacks, as the fake class could not be well represented by training examples. We proposed a **novel one-class learning approach** to address this challenge [26] (Figure 3). We assume that the bona fide class is well represented by training utterances, but the fake class is not. This approach learns an embedding space such that bona fide utterances are compacted to a small region while fake utterances are pushed away from that region, significantly improving the generalization ability to unseen attacks. Since its publication in 2021, our paper [26] has been **cited 257 times** according to Google Scholar.

In the last two years, we extended our deepfake detection to singing voices and audiovisual data. Singing voice deepfakes are becoming increasing concerns of singers and music labels. We **proposed the singing voice deepfake detection (SVDD)** task [27] and **organized the inaugural SVDD challenge** at the 2024 IEEE Workshop on Spoken Language Technology (SLT) [28, 29] together with collaborators from CMU and Nagoya University. Video deepfakes are becoming significant concerns on social media. We extended the one-class learning idea to **audiovisual deepfake detection**, i.e., using both audio and visual modalities to detect deepfake video [30].

On the generation side, we have been working on **controllable, expressive** and **conversational** speech synthesis, which I argue are the new frontiers of speech synthesis research. On voice conversion, i.e., converting one person's voice to another person without changing the linguistic content, we **proposed the first controllable voice conversion method that allows users to apply time-varying controls on pitch and speed** of the utterance [31]. It integrates signal processing techniques (i.e., TD-PSOLA) and deep neural networks toward fine-grained controls on human interpretable parameters (i.e., pitch and speed). On speech expressiveness modeling, we **extended the definition of expressiveness from the commonly modeled high-level aspects (e.g., emotion and style) to articulatory-level aspects** including Glottalization, Tenseness and Resonance (GTR) [32]. This was inspired by the professional practices of voice acting, where expressiveness is needed at both levels which are often independent from each other. We recorded the GTR-Voice dataset containing utterances spoken by one professional voice actor with 125 combinations of GTR configurations and investigated text-to-speech approaches with controls on the GTR dimensions. On conversional speech synthesis, we trained a Transformer-based model named Parakeet on 100,000 hours of conversational data including Spotify Podcasts, LibriVox, and Common Voice [33]. We finetuned the Whisper model (i.e., a pre-trained automatic speech transcription model from OpenAI) to transcribe the conversations, applied the Descript Audio Codec (DAC)

to encode conversation audio, and trained an autoregressive transformer with classifier-free guidance (CFG) to generate audio from conversational text input. Parakeet achieves **extraordinary naturalness of conversational TTS**, surpassing all existing models that I have seen. This model has been used to generate conversational training data to improve conversational ASR [34]. However, due to the high societal risks, we choose not to open-source the model at this moment.

For future work, we are working on real-time speech anonymization with static trait (e.g., gender, age, dialect) control and dynamic trait (e.g., emotion) normalization for the purpose of speaker privacy protection. This project is decomposed into three main modules: 1) pseudo-speaker generation, 2) streaming voice morphing, 3) speaker trait disentanglement, and 4) dynamic trait normalization. We have been applying and developing different generative AI techniques including flow models, generative adversarial networks (GANs), and diffusion models toward these tasks. On diffusion models for audio applications, we are building on our previous work on diffusion-based sound effect generation [35] to develop a code repository that covers various applications including speech enhancement, source separation, text-to-speech synthesis, and text-to-audio generation.

Personalized and Adaptive Spatial Audio Rendering

Spatial audio is critical in creating immersive experiences in multimedia content delivery. It has many applications in entertainment, education, healthcare and communication. Spatial audio is also an emerging area where acoustics, signal processing and machine learning intersects. In the past few years, I started working on this direction with my collaborators, with the support from Goergen Institute for Data Science seed funding ($50,000, 01/2023-12/2023 and 01/2025-12/2025, PI).

There are two primary spatial audio rendering approaches: *headphone-based* and *loudspeaker-based*. The headphone-based approach applies head-related transfer functions (HRTF) to sound source signals. HRTFs describe the modification of sound due to ear and head geometry of the listener when sound travels from the source to the listener's ears. Therefore, HRTFs are different for different people; applying the wrong HRTFs would cause localization confusion for the listener [36]. However, measuring HRTFs for a listener is very time-consuming (e.g., 2 hours) and requires special equipment (e.g., anechoic chamber, rotatable loudspeaker array). There is a strong need for predicting HRTFs for each listener by analyzing their ear and head geometry. Classical approaches such as the boundary element method (BEM) are computationally expensive and show significant deviations between the ground-truth measurements and the predicted HRTFs [37]. Machine learning based methods are promising, but they suffer from the small dataset sizes and different spatial sampling schemes that different datasets use.

We **proposed to learn a unified representation of HRTFs regardless spatial sampling schemes**. This is a neural field representation and is named HRTF Field [38, 39]. A neural field is a latent, unified, and differentiable representation learned by a deep neural network from discrete samples of a function defined in a continuous space. This representation is unified across datasets with different spatial sampling schemes, allowing HRTFs for arbitrary azimuth and elevation angles to be represented. This approach shows significantly better performance on HRTF interpolation and generation than existing approaches. This work is recognized as **among the top 3% of accepted papers** at ICASSP 2023.

On loudspeaker-based rendering, a key challenge is the control of room acoustics, which describes the reflections of sounds at room surfaces before they reach listeners' ears. Rooms of different geometries and surface materials can have drastically different acoustics (e.g., church vs. living room). Without appropriate control of room acoustics, loudspeaker-based rendering is very much limited. Along this direction, we are working on an **innovative approach toward active control of room acoustics**. Our idea is to use flat-panel loudspeakers and develop algorithms to control panel reflections to incoming sound waves in real time. The flat-panel loudspeaker technology is a transformative innovation by my colleagues at UofR in recent years. Through vibration actuators and sensors behind, the panel can function as a loudspeaker and microphone at the same time. If one could predict the incoming wave to a panel, then a cancelation signal could be prepared to cancel the reflection at the panel in real time. This would require an adaptive feedback control algorithm (signal processing) and a waveform prediction algorithm (generative machine learning). This approach could also allow for separate reflection controls for different types of sound sources (e.g., music, speech) and sources at different locations. This highly innovation project is being funded by the Goergen Institute for Data Science seed funding program.

References

[1] Y. Yan, F. Cwitkowitz, and Z. Duan, "Skipping the Frame-Level: Event-Based Piano Transcription With Neural Semi-CRFs," *Advances in Neural Information Processing Systems,* vol. 34, pp. 20583-20595, 2021.

[2] Y. Yan and Z. Duan, "Scoring Intervals using Non-hierarchical Transformer For Automatic Piano Transcription," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[3] F. Cwitkowitz, J. Driedger, and Z. Duan, "A Data-Driven Methodology for Considering Feasibility and Pairwise Likelihood in Deep Learning Based Guitar Tablature Transcription Systems," in *Sound and Music Computing Conference*, 2022.

[4] Y. Zang, Y. Zhong, F. Cwitkowitz, and Z. Duan, "SynthTab: Leveraging Synthesized Data for Guitar Tablature Transcription," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[5] H. Yu and Z. Duan, "Note-level transcription of choral music," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[6] F. Cwitkowitz and Z. Duan, "Toward Fully Self-Supervised Multi-Pitch Estimation," *arXiv preprint arXiv:2402.15569,* 2024.

[7] M. Heydari, J.-C. Wang, and Z. Duan, "SingNet: a real-time Singing Voice beat and Downbeat Tracking System," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: IEEE, pp. 1-5.

[8] M. Heydari, M. McCallum, A. Ehmann, and Z. Duan, "A Novel 1D State Space for Efficient Music Rhythmic Analysis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022: IEEE, pp. 421-425.

[9] M. Heydari and Z. Duan, "Don't Look Back: An Online Beat Tracking Method Using RNN and Enhanced Particle Filtering," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021: IEEE, pp. 236-240.

[10] M. Heydari, F. Cwitkowitz, and Z. Duan, "BeatNet: A real-time music integrated beat and downbeat tracker," *International Society for Music Information Retrieval,* 2021.

[11] M. Heydari and Z. Duan, "BeatNet+: Real-Time Rhythm Analysis for Diverse Music Audio," *Transactions of the International Society for Music Information Retrieval,* vol. 7, no. 1, 2024.

[12] Y. Yan, E. Lustig, J. VanderStel, and Z. Duan, "Part-invariant Model for Music Generation and Harmonization," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 204-210.

[13] C. Benetatos, J. VanderStel, and Z. Duan, "BachDuet: A deep learning system for human-machine counterpoint improvisation," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2020.

[14] N. Jiang, S. Jin, Z. Duan, and C. Zhang, "RL-Duet: Online Music Accompaniment Generation Using Deep Reinforcement Learning," in *AAAI*, 2020.

[15] N. Jiang, S. Jin, Z. Duan, and C. Zhang, "When counterpoint meets Chinese folk melodies," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, vol. 33, pp. 16258-16270.

[16] G. Zhu, J.-P. Caceres, Z. Duan, and N. J. Bryan, "MusicHiFi: Fast high-fidelity stereo vocoding," *IEEE Signal Processing Letters,* 2024.

[17] C. Benetatos and Z. Duan, "Draw and Listen! A Sketch-Based System for Music Inpainting," *Transactions of the International Society for Music Information Retrieval,* vol. 29, pp. 2288-2292, 2022.

[18] H. F. Garcia, C. Benetatos, P. O'Reilly, A. Aguilar, Z. Duan, and B. Pardo, "HARP: Bringing Deep Learning to the DAW with Hosted, Asynchronous, Remote Processing," 2023: NeurIPS 2023 Workshop on Machine Learning for Creativity and Design.

[19] C. C. Benetatos, Frank; Pruyne, Nathan; García, Hugo Flores; O'Reilly, Patrick; Duan, Zhiyao; Pardo, Bryan, "HARP 2.0: Expanding Hosted, Asynchronous, Remote Processing for Deep Learning in the DAW," presented at the International Society for Music Information Retrieval Conference (ISMIR) Late Breaking & Demo, 2024.

[20] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Int. Conference on Digital Audio Effects (DAFx-10)*, 2010.

[21] H. F. Garcia, P. Seetharaman, R. Kumar, and B. Pardo, "Vampnet: Music generation via masked acoustic token modeling," *arXiv preprint arXiv:2307.04686,* 2023.

[22] F. Cwitkowitz *et al.*, "Timbre-Trap: A Low-Resource Framework for Instrument-Agnostic Music Transcription," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024: IEEE, pp. 1291-1295.

[23] C.-Z. A. Huang *et al.*, "Music transformer," *arXiv preprint arXiv:1809.04281,* 2018.

[24] L. Zhang, X. Wang, E. Cooper, N. Evans, and J. Yamagishi, "The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 31, pp. 813-825, 2022.

[25] C. Donahue, I. Simon, and S. Dieleman, "Piano genie," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 160-164.

[26] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters,* vol. 28, pp. 937-941, 2021.

[27] Y. Zang, Y. Zhang, M. Heydari, and Z. Duan, "SingFake: Singing Voice Deepfake Detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[28] Y. Zang *et al.*, "CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection," in *Interspeech*, 2024.

[29] Y. Zhang, Y. Zang, J. Shi, R. Yamamoto, T. Toda, and Z. Duan, "SVDD 2024: The Inaugural Singing Voice Deepfake Detection Challenge," in *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

[30] K. Lee, Y. Zhang, and Z. Duarr, "A Multi-Stream Fusion Approach with One-Class Learning for Audio-Visual Deepfake Detection," in *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, 2024: IEEE, pp. 1-6.

[31] M. Chen and Z. Duan, "ControlVC: Zero-Shot Voice Conversion with Time-Varying Controls on Pitch and Rhythm," in *Interspeech*, 2023.

[32] Z. K. Li, M. M. Chen, Y. Zhong, P. Liu, and Z. Duan, "GTR-Voice: Articulatory Phonetics Informed Controllable Expressive Speech Synthesis," in *Interspeech*, 2024, vol. 2024, pp. 1775-1779.

[33] J. Z. Darefsky, Ge; Duan, Zhiyao, "Parakeet: A natural sounding, conversational text-to-speech model," vol. 2025, ed, 2024.

[34] S. Cornell, J. Darefsky, Z. Duan, and S. Watanabe, "Generating data with text-to-speech and large-language models for conversational speech recognition," presented at the Interspeech Workshop on SynData4GenAI, 2024.

[35] G. Zhu, Y. Wen, M.-A. Carbonneau, and Z. Duan, "EDMSound: Spectrogram Based Diffusion Models for Efficient and High-Quality Audio Synthesis," *NeurIPS Workshop on Machine Learning for Audio,* 2023.

[36] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database," *Applied Sciences,* vol. 8, no. 11, p. 2029, 2018.

[37] Z. Stanford, "PERSONALIZED HRTF MODELING USING DNN-AUGMENTED BEM," *Measurement,* vol. 200, p. 300.

[38] Y. Zhang, Y. Wang, and Z. Duan, "HRTF Field: Unifying Measured HRTF Magnitude Representation with Neural Fields," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023: IEEE, pp. 1-5.

[39] Y. Wen, Y. Zhang, and Z. Duan, "Mitigating Cross-Database Differences for Learning Unified HRTF Representation," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023: IEEE, pp. 1-5.