Li *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2024) 2024:44
https://doi.org/10.1186/s13636-024-00353-7

EURASIP Journal on Audio, Speech, and Music Processing

EDITORIAL

Open Access

Guest editorial: AI for computational audition—sound and music processing

Zijin Li[1*], Wenwu Wang[2], Kejun Zhang[3] and Mengyao Zhu[4]

Abstract

Nowadays, the application of artificial intelligence (AI) algorithms and techniques is ubiquitous and transversal. Fields that take advantage of AI advances include sound and music processing. The advances in interdisciplinary research potentially yield new insights that may further advance the AI methods in this field. This special issue aims to report recent progress and spur new research lines in AI-driven sound and music processing, especially within interdisciplinary research scenarios.

1 Introduction

Despite the long history in the development of AI technologies, their applications in audio are still in the early stage, and the user experience of related audio products is far from satisfactory. There also exists a gap between the current generation of audio technologies and those that will be needed for future interactive applications. In addition, how to balance model performance and available resources in practical applications, such as computing, storage, and transmission, is one of the main problems faced by audio-intelligent computing systems. In some applications with real-time requirements, intelligent scheduling of network resources is particularly important.

This Special Issue aims to collect research on AI for computational audition including music, speech, and general sound. The principal goal is to bring together scholars interested in the research on the theory and technology to realize the integration of traditional methods and emerging technologies, the application and comparative analysis of different intelligent technologies in music creation, audio processing and detection, and recognition. This issue has accepted a total of 11 relevant articles, primarily categorized into three thematic areas: "AI for the recognition and analysis of music," "AI for speech processing and applications," and "AI for general audio and sound." These themes are formed based on the research presented in the articles.

2 AI for the recognition and analysis of music

In the eleven papers, five focus on music features, distinctions, recognition, and generation utilizing AI technologies. These works showcase AI's impact across various facets of music, ranging from using traditional Chinese music's aesthetic features to enhance synthesized guzheng music quality, addressing melody-harmony relations, improving video-music retrieval, to resolving bandwidth extension in music signals.

The paper titled *Acoustical Feature Analysis and Optimization for Aesthetic Recognition of Chinese Traditional Music* authored by Lingyun Xie, Yuehong Wang, and Yan Gao focuses on researching the aesthetic characteristics of traditional Chinese music. The paper begins by introducing a database containing various forms of traditional Chinese music. Employing two feature selection methods (filtering and wrapping), the authors extract 44 dimensions of features suitable for the aesthetic classification of traditional Chinese music from a set of 447 low-level features. Through these features, the paper further investigates the correlation between musical elements and

*Correspondence:
Zijin Li
lzijin@ccom.edu.cn
[1] Central Conservatory of Music, Beijing, China
[2] University of Surrey, Guildford, UK
[3] Zhejiang University, Hangzhou, China
[4] Shanghai University, Shanghai, China

aesthetic classification. Detailed analyses and studies of different aesthetic categories within traditional Chinese music are presented, offering valuable references for further studies in information retrieval and intelligent processing of music.

The paper titled *Effective Acoustic Parameters for Automatic Classification of Performed and Synthesized Guzheng Music* authored by Huiwen Xue, Chenxin Sun, Mingcheng Tang, Chenrui Hu, Zhengqing Yuan, Min Huang, and Zhongzhe Xiao explores the acoustic differences between synthesized and real guzheng (a traditional Chinese musical instrument) performances with the aim of enhancing the quality of synthesized guzheng music. This study, based on a dataset from various sources and genres, analyzes the automatic classification problem of guzheng music. The results indicate high classification accuracy (93.30%) with a single feature, demonstrating significant differences between synthesized and real guzheng music. Through the combination of complementary features, the study achieves near-perfect classification accuracy (99.73%). The conclusion suggests that future improvements in guzheng synthesis algorithms may involve incorporating spectral flow attributes. This research not only deepens the understanding of guzheng timbre analysis but also provides new research directions for audio synthesis techniques for traditional music instruments.

In the paper titled *Generating Chord Progression from Melody with Flexible Harmonic Rhythm and Controllable Harmonic Density*, the authors Shangda Wu, Yue Yang, Zhaowen Wang, Xiaobing Li, and Maosong Sun present a new system named AutoHarmonizer, designed to address melody and harmonization issues, specifically how to generate chord progressions for a given melody. AutoHarmonizer emphasizes controllable harmonic rhythm and harmony density, boasting an extensive lexicon of 1462 chord types. The system's key innovation lies in its ability to flexibly adjust harmonic rhythm, aiming to create expressive and musically logical harmonic effects. This system holds potential applications in music composition, arrangement, and production, providing an automated tool for harmonic composition. Despite successful chord progression generation using neural networks in previous research, limitations exist in controlled melody and harmonization. AutoHarmonizer fills this gap, and experimental results demonstrate its diverse harmonic rhythm and effective controllable harmony density.

The paper titled *YuYin: A Multi-task Learning Model of Multi-modal E-commerce Background Music Recommendation* authored by Le Ma, Xinda Wu, Ruiyuan Tang, Chongjun Zhong, and Kejun Zhang investigates applications in music video recognition and retrieval. The authors initially established a large-scale e-commerce advertising dataset, "Commercial-98 K". Subsequently, they propose a video-music retrieval model, "YuYin," designed to learn the association between videos and music. The model integrates emotional and audio features of music through a Weighted Fusion Module (WFM) to obtain a more detailed music representation. Considering the similarity of music within the same product category, "YuYin" is trained through multi-task learning, exploring the association between video and music through cross-matching tasks involving video, music, labels, and category predictions. Through extensive experiments, the authors demonstrate the significant improvement of "YuYin" in video-music retrieval on the "Commercial-98 K" dataset.

The paper titled *Efficient Bandwidth Extension of Musical Signals Using a Differentiable Harmonic Plus Noise Model* authored by Pierre-Amaury Grumiaux and Mathieu Lagrange primarily addresses the bandwidth extension issue in music signals using a differentiable digital signal processing (DDSP) model. This model, incorporating neural networks, is trained to infer parameters for the digital signal processing model, effectively generating full-bandwidth audio signals. The research initially focuses on bandwidth extension for monophonic signals and proposes two methods for handling multichannel signals. Evaluations performed on synthesized monophonic and multichannel data, compared with baseline models and state-of-the-art deep learning models, indicate the superiority of the proposed model in objective frequency-domain metrics. Furthermore, the authors evaluate these models on real-world data, encompassing monophonic and multichannel scenarios with various instruments and music types. Through the MUSHRA listening tests, they further confirm the superiority of the proposed methods.

3 AI for speech processing and applications

The next three papers primarily focus on speech, including an artificial intelligence model for speech separation and a humorous speech database for driving environments. AI's influence on speech processing is exemplified in the following three papers, including speech separation, lightweight speaker separation, and the impact of humorous speech on driver emotions in congested traffic.

In the paper titled *Deep Encoder/Decoder Dual-Path Neural Network for Speech Separation in Noisy Reverberation Environments*, the authors Chunxi Wang, Maoshen Jia, and Xinfeng Zhang propose a novel speech separation model and conduct subjective and objective experiments to compare its performance with other reference methods. The experiments indicate superior separation performance in complex acoustic environments. In addition, the authors include experiments in real-world conditions to further validate the model's practical applicability. The paper concludes with supplementary enhancements based on reviewers' suggestions and expresses the intention to further improve the model and test its generalization on broader datasets.

In the paper titled *Lightweight Target Speaker Separation Network Based on Joint Training*, the authors Jing Wang, Hanyue Liu, Liang Xu, Wenjing Yang, Weiming Yi, and Fang Liu introduce a lightweight target speaker separation network based on Long Short-Term Memory (LSTM) networks. This aims to address the issues of system latency and performance limits resulting from the large model size in existing deep learning separation methods. The network achieves a reduction in model size and computation latency while maintaining satisfactory separation performance through optimized network structure and training methods. The authors propose a target speaker separation method based on joint training, utilizing a joint loss function (speaker registration and separation) to achieve overall training and optimization of the target speaker separation system. Experimental results demonstrate that this lightweight network outperforms the original model in both size reduction and separation performance. The introduced joint training loss function further enhances separation performance.

In the paper titled *The Power of Humorous Audio: Exploring Emotion Regulation in Traffic Congestion through EEG-based Study*, the authors Lekai Zhang, Yingfan Wang, Kailun He, Hailong Zhang, Baixi Xing, Xiaofeng Liu, and Fo Hu investigate the regulatory effect of humorous language on driver anger emotions under congested traffic conditions. The study employs 50 samples of humorous speech, rated high on four to five measurement dimensions. Using a comparative approach combined with subjective emotion assessments and electroencephalogram (EEG) data, the study evaluates the regulatory effect of humorous speech on the emotional state of drivers. The research also suggests potential future directions, including comparisons with the emotional regulation effects of positive music and the analysis of specific regulatory mechanisms of humorous speech. This study provides new insights for the design of road rage management systems.

4 AI for general audio and sound

The scope of artificial intelligence in audio processing extends beyond music and speech; it also encompasses the broader recognition, processing, and analysis of general sounds. There are three papers exploring distinct aspects including soundscape reconstruction utilizing Neural Processes and Dynamic Cores, unsupervised anomaly sound detection employing a Transformer-based Autoencoder, and identifying snoring sounds under limited data resources using Meta-Learning techniques.

In the paper titled *Sound Field Reconstruction Using Neural Processes with Dynamic Kernels*, the authors Zining Liang, Wen Zhang, and Thushara D. Abhayapala employ advanced techniques involving Neural Processes (NPs) and dynamic cores to reconstruct soundscapes. Traditional methods utilize Gaussian Processes (GPs) and fixed cores to model spatial correlations in soundscapes, but they have limitations such as limited expressive power of cores and the need for manual identification of optimal cores for different soundscapes. The study introduces a novel approach that utilizes deep neural networks based on NPs to parameterize GPs, enabling the dynamic learning of cores from simulated data. The incorporation of attention mechanisms enhances the flexibility and adaptability of the proposed method to the acoustic properties of soundscapes. Numerical experiments demonstrate that this new method outperforms existing approaches in terms of reconstruction accuracy, offering a promising alternative for soundscape reconstruction.

In the paper titled *Transformer-based Autoencoder with ID Constraint for Unsupervised Anomalous Sound Detection*, the authors Jian Guan, Youde Liu, Qiuqiang Kong, Feiyang Xiao M.D., Qiaoxi Zhu, Jiantong Tian, and Wenwu Wang propose a Transformer-based autoencoder architecture, IDC-TransAE, for unsupervised anomaly sound detection (ASD). This method leverages machine IDs to constrain the latent space of the autoencoder and introduces a simple ID classifier to learn distribution differences among the same machine types, thereby enhancing the model's capability to distinguish abnormal sounds. The authors also introduce a method for computing weighted anomaly score to highlight the anomaly scores of events that occur only briefly. Experimental results on the DCASE 2020 Challenge Task 2 development dataset demonstrate the effectiveness and superiority of this method. Overall, this research provides a new methodological approach for detecting unknown anomalous sounds in devices using only normal sound data.

The paper titled *Battling with the Low-Resource Condition for Snore Sound Recognition: Introducing a Meta-Learning Strategy*, authored by Jingtan Li, Mengkai Sun, Zhonghao Zhao, Xingcan Li, Gaigai Li, Chen Wu, Kun Qian, Bin Hu, Yoshiharu Yamamoto, and Björn W. Schuller focuses on the recognition of snoring sounds under limited resources. The study effectively addresses the challenge of limited sample data by employing a few-shot learning method called "Model-Agnostic Meta-Learning (MAML)." The research achieves a significant unweighted average recall rate of 60.2% on the test dataset. This work significantly contributes to the diagnosis and treatment of diseases such as obstructive sleep apnea. The primary contribution of this study lies in the application of meta-learning strategies to enhance the recognition efficiency of snoring sounds under resource constraints.

5 Conclusion

The special issue illustrates the extensive application of artificial intelligence to numerous problems in music, speech, and general audio processing. These studies demonstrate the potential of interdisciplinary research to enhance AI methodologies in this field while identifying the existing gap between current technology and future requirements. They offer valuable insights for future research, showcasing the potential of various novel techniques across diverse domains and establishing a solid foundation for advancements in these areas. With ongoing technological advancements and continuous innovation, artificial intelligence will continue to play a crucial role in audio processing, nurturing new prospects and possibilities for music, speech, and sound processing.

Acknowledgements
Not applicable.

Authors' contributions
Not applicable.

Funding
Not applicable.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 7 May 2024 Accepted: 14 May 2024
Published online: 11 September 2024

Publisher's Note

Zijin Li is a Professor with the Department of AI Music and Music Information Technology, Central Conservatory of Music. She received the Ph.D. degree in music acoustics from Central Conservatory of Music in 2013. She once was a Visiting Professor at McGill University, Canada in 2019 and a Senior Visiting Professor at Tsinghua University, China in 2021. Her current research interests include music acoustics, music creativity, new musical instrument design, and Innovation theory of music technology. She served as chief guest editor of teh Journal of Cognitive Computation and Systems (JCCS) Special Issue: Perception and Cognition in Music Technology and guest editor of Frontiers: Human-Centred Computer Audition: Sound, Music, and Healthcare.

Wenwu Wang is a Professor in Signal Processing and Machine Learning, and a Co-Director of the Machine Audition Lab within the Centre for Vision Speech and Signal Processing, University of Surrey, UK. He is also an AI Fellow at the Surrey Institute for People-Centred Artificial Intelligence. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 300 papers in these areas. His research has been funded by the UK and EU research councils and leading international companies (including Samsung, Tencent, Huawei, BBC, and Saab), as principal investigator or co-investigator on over 30 projects with a total award value of over £30 million. He is a (co-)author or (co-)recipient of over 15 awards including the IEEE Signal Processing Society 2022 Young Author Best Paper Award, Best Paper Award on ICAUS 2021, Judge's Award on DCASE 2020 and DCASE 2023, the Reproducible System Award on DCASE 2019 and 2020, Best Student Paper Award on LVA/ICA 2018, the Best Oral Presentation on FSDM 2016, Best Student Paper Award finalists on ICASSP 2019 and LVA/ICA 2010, the TVB Europe Award for Best Achievement in Sound in 2016, and the Best Solution Award on the Dstl Challenge in 2012. He is on the Stanford University List of Top 2% Scientists Worldwide during 2021–2023.

He is an Associate Editor (2020–2025) for IEEE/ACM Transactions on Audio Speech and Language Processing, an Associate Editor for (Nature) Scientific Report, a Specialty Editor in Chief of Frontier in Signal Processing, an Area Editor for, an Associate Editor for EURASIP Journal on Audio Speech and Music Processing. He was a Senior Area Editor (2019–2023) and an Associate Editor (2014–2018) for IEEE Transactions on Signal Processing. He is the Elected Chair (2023–2024) of the IEEE Machine Learning for Signal Processing Technical Committee, the elected Vice Chair (2022–2024) of the EURASIP Technical Area Committee for Audio Speech and Music Processing, an elected Member (2021–2023) of the IEEE Signal Processing Theory and Methods Technical Committee, and an elected Member of the International Steering Committee of Latent Variable Analysis and Signal Separation. He was a Satellite Workshop Co-Chair for INTERSPEECH 2022, Publication Co-Chair for IEEE ICASSP 2019, Local Arrangement Co-Chair of IEEE MLSP 2013, and Publicity Co-Chair of IEEE SSP 2009. He is a Satellite Workshop Co-Chair for ICASSP 2024, and Special Session Co-Chair for MLSP 2024.

Kejun Zhang is a Professor of Zhejiang University, Joint PhD supervisor on Design and Computer Science, and Director of Next Lab of Zhejiang University. He received his PhD degree from the College of Computer Science and Technology, Zhejiang University in 2010. From 2008 to 2009, He was a visiting research scholar of the University of Illinois at Urbana-Champaign, USA. After a multidisciplinary academic education, he worked as a Post-doctor at Zhejiang University for 3 years and at MONASH University for nearly half years. In June 2013, he became a faculty of the College of Computer Science and Technology at Zhejiang University.

His research interests are Design Science, Artificial Intelligence, Multimedia Computing and Design, and the understanding, modeling, and innovation design of products and social management by computational means. He is now the PI of the National Science Foundation of China, Co-PI of the National Key Research and Development Program of China, and PIs of ten more other research programs. He has authored 2 books and more than 40 scientific papers.

Mengyao Zhu is a technical expert in the Audio Department of Huawei CBG since 2019 on sabbatical leave from Shanghai University. He was an Associate Prof. of School of Communication and Information Engineering, Shanghai University, since 2011. He received the B.S. and the Ph.D. degree in Communication and Information System from Zhejiang University, Hangzhou, China, in 2004 and 2009, respectively. He is now in charge the feature of spatial audio on consumer device of Huawei. He was serving as TPC Co-Chair of Conference on Sound and Music Technology (CSMT) in 2020 and 2021. His research interests include sound field capture and reproduction, audio and speech signal processing, and circuits and system design of multimedia system.