Speech Removal Framework for Privacy-preserving Audio Recordings

Gabriel Bibbó, Arshdeep Singh, Thomas Deacon, Mark D. Plumbley

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK,
{g.bibbo, arshdeep.singh, m.plumbley}@surrey.ac.uk

***Abstract*—Public dataset such as "The Sounds of Home" [1] are being recorded in people's home to capture everyday soundscape. Such audio recordings from home environments provide valuable information for recognizing daily activities, monitoring health and wellbeing, and enabling smart home applications. They support the development of robust sound event detection systems under real-world conditions. However, in-home recordings contain crucial personal information in the form of speech signals. It is crucial to remove the personal information such as "speech" from domestic audio recordings when publicly sharing the recorded datasets. This demonstration showcase real-time identification of personal information, in our case it is speech, using various AI models such as convolutional neural networks (PANNs, E-PANNs), Transformer model (AST), voice activity detection (VAD) models (Silero, WebRTC). Our focus is two fold: (1) To design a speech removal system to identify and remove speech from the recorded audio in real-time. (2) How well can AI models distinguish speech from non-speech audio? Our demonstration is simple, easy to use and a software-based GUI.**

1. INTRODUCTION

With the rise of smart home technologies, there is growing potential to improve the quality of life for older adults through the use of intelligent audio systems. Advances in artificial intelligence (AI) are enabling new ways to utilize sound for in-home wellbeing—ranging from acoustic environment monitoring to interactive and responsive audio-based applications [2]. A key challenge facing audio-based AI systems is the requirement for large volumes of task-specific data to ensure reliable performance [3]. Recently, datasets such as "The Sounds of Home [1]" which is a large-scale (1300+ hours) publicly available dataset, recorded within homes, to support the design and deployment of AI sound technologies aimed at improving quality of life for older adults [4].

However, in-home recordings introduces significant privacy and data governance concerns as they have speech recordings, particularly in light of regulations such as General Data Protection Regulation (GDPR) [5].

To address the privacy related challenges while making available recorded dataset publicly, our demonstration presents an AI based speech anonymization framework which can identify and remove speech form the audio recordings before they made publicly available. Our demonstration can be utilized to automatically remove speech from large-scale recordings and can also run in real-time with some delay while storing the recordings into buffer. We benchmark our speech annoymization framework on publicly available CHiME home dataset [6] using various AI models including Transformer [7] and Convolutional Neural Networks (CNNs) [8], [9].

We aim for our work to highlight both existing and emerging challenges faced by the Detection and Classification of Acoustic Scenes and Events (DCASE) community, particularly in addressing the removal of personal information from audio recordings. Despite current limitations, we present our approach as a practical framework for deploying such systems on embedded devices, supporting their wider adoption in real-world applications.

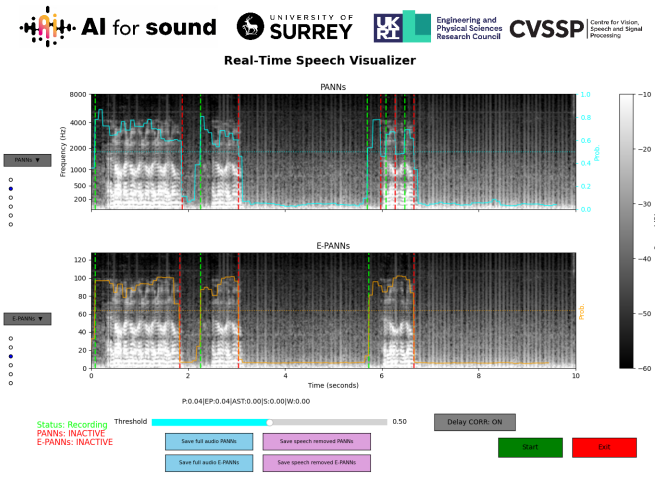

Fig. 1: Real-time speech detection interface for analysing the performance across different AI models.

2. SPEECH REMOVAL FRAMEWORK

Our speech removal framework is simple and consists of three steps. First step includes recording the sounds or uploading the already recorded audio clips. In the second step, a pre-trained audio AI model is used to detect the speech-related event in the input recording. Finally, the speech-related chunk is replaced with the silence from the audio is removed from the input recording.

2.1. Functionality of Speech Removal Demonstration

Our demonstration shows a real-time speech detection and removal system with live spectrogram visualization. The graphical user interface (GUI) lets researchers compare different AI models for detecting speech in domestic audio environments and export privacy-preserved recordings 1.

Core Features:

- **Real-time Audio Processing:** The system captures audio at 32 kHz through a ring buffer, processing 4-second chunks continuously. A 10-second circular buffer stores recent audio for export.
- **Multi-Model Support:** Six AI models work together: PANNs (CNN14 with decision-level attention) [8], E-PANNs (uses 73% fewer parameters) [10], AST (Audio Spectrogram Transformer) [7], Silero-VAD (lightweight neural VAD) [11], and WebRTC-VAD (classic signal processing baseline). Users can select two models at once for side-by-side comparison.
- **Live Visualization:** Two mel-spectrogram displays show 10 seconds of audio history with 128 mel bins (20-8000 Hz). Speech probability curves overlay the spectrograms with model-specific colors. Green markers show speech onset, red markers show offset. Users adjust detection thresholds (0-1) with a slider.

- **Dynamic Delay Calibration:** The system can measure and correct for model-specific processing delays, aligning detection events with the actual audio signal.
- **Privacy-Preserving Export:** Two buttons provide practical options: "Save full audio" exports the 10-second buffer with JSON annotations containing detection timestamps. "Save speech removed" exports audio with detected speech replaced by silence, protecting privacy while keeping non-speech sounds.

Users press "Start" to begin recording, then speak or play audio samples. The system shows detection results in real-time across selected models. When users press either save button, the system creates timestamped WAV files and JSON files with precise speech boundaries from each model and the threshold used. This supports both privacy protection and reproducible analysis.

3. BENCHMARKING OF SPEECH REMOVAL FRAMEWORK ON CHIME-HOME DATASET

We evaluated our speech detection framework on the CHiME-Home dataset [6], where only clips containing child (`c`), male (`m`), or female (`f`) speech are considered positive. This represents the core task of detecting human vocalizations in domestic environments.

3.1. Performance Analysis

Table 1 presents the comparative performance of five AI models integrated into our demonstration system. We report F1-score at optimal thresholds, model parameters, and Real-Time Factor (RTF) measured on 4-second audio chunks.

Table 1: Performance comparison of speech detection models on CHiME-Home

Model	F1-Score	Parameters	RTF
AST	0.860	88M	0.039
PANNs	0.848	81M	1.263
E-PANNs	0.847	22M	0.154
Silero-VAD	0.806	1.8M	0.057
WebRTC-VAD	0.708	$<$0.1M	0.002

The results reveal important trade-offs between accuracy and computational efficiency:

Accuracy Leaders: AST achieves the highest F1-score (0.860), followed closely by PANNs (0.848) and E-PANNs (0.847). These large pre-trained models demonstrate superior ability to distinguish speech from complex domestic sounds.

Efficiency Champions: WebRTC-VAD offers exceptional speed (RTF=0.002) suitable for extremely resource-constrained deployments, though with reduced accuracy. Silero-VAD provides an attractive middle ground with competitive performance (F1=0.806) and low computational cost (RTF=0.057).

Practical Considerations: PANNs exceeds real-time constraints (RTF$>$1.0), limiting its deployment on edge devices. E-PANNs successfully reduces computational cost by 88% while maintaining comparable accuracy, demonstrating the effectiveness of model compression techniques.

These benchmarks, combined with our interactive demonstration, enable practitioners to make informed decisions when selecting speech detection models for privacy-preserving applications, balancing accuracy requirements against available computational resources.

4. LIST OF EQUIPMENT AND PHYSICAL SPACE

We need a power socket and a table for laptop.

5. CONCLUSIONS

This demonstration provides real-time analysis of speech detection models for privacy-sensitive audio scenarios. Users can select models, adjust detection thresholds, and watch speech detection states and probabilities change in real-time. Our benchmarking on CHiME-Home shows that model selection must balance performance needs with computational limits on edge devices.

Our tool solves the challenge of privacy protection in domestic audio research. Users can export both annotated recordings and privacy-preserved versions (with speech removed), helping researchers build realistic datasets while following privacy regulations like GDPR. The side-by-side model comparison and dynamic delay calibration reveal practical deployment challenges often missed in academic evaluations.

Our results show that transformer-based AST provides the best balance for edge deployment (F1=0.860, RTF=0.039), while Silero-VAD works well for highly constrained environments. The JSON annotations support reproducibility and help analyze when models disagree, which matters for building robust privacy-preserving pipelines. This demonstration bridges the gap between theoretical model performance and real-world deployment needs, supporting privacy-by-design audio systems for practical applications.

6. ACKNOWLEDGMENT

This work was supported by Engineering and Physical Sciences Research Council (EPSRC) Grant EP/T019751/1 "AI for Sound (AI4S)". For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

REFERENCES

[1] G. Bibbó, T. Deacon, A. Singh, and M. D. Plumbley, "The Sounds of Home: A Speech-Removed Residential Audio Dataset for Sound Event Detection," in *Proc. CHiME 2024*, pp. 49–53.

[2] E. Corrigan-Kavanagh, A. Fernandez, and M. Plumbley, "Envisioning sound sensing technology for enhancing urban living," in *Environments By Design*. Architecture Media Politics Society, Jan. 2022, p. 11.

[3] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, Sept. 2021.

[4] T. Deacon, D. Frohlich, M. D. Plumbley, G. Bibbo, and A. Singh, "Sound wellbeing in later life," University of Surrey, 2023, (accessed on June 26, 2024). [Online]. Available: https://www.surrey.ac.uk/digital-world-research-centre/funded-projects/sound-wellbeing-later-life

[5] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," Official Journal of the European Union, 2016, (accessed on June 18, 2024). [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[6] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "CHiME-home: A dataset for sound source recognition in a domestic environment," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[7] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[8] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[9] A. Singh and M. D. Plumbley, "Efficient CNNs via Passive Filter Pruning," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 1763–1774, 2025.

[10] A. Singh, H. Liu, and M. D. Plumbley, "E-PANNs: sound recognition using efficient pre-trained audio neural networks," in *Inter-Noise and Noise-Con Congress and Conference Proceedings*, vol. 268, no. 1. Institute of Noise Control Engineering, 2023, pp. 7220–7228.

[11] Silero AI Team, "Silero vad: Pre-trained enterprise-grade voice activity detector," GitHub repository, https://github.com/snakers4/silero-vad, 2024, mIT License; accessed 2025-07-07.