PRIVACY-PRESERVING OCCUPANCY ESTIMATION

Jennifer Williams, Vahid Yazdanpanah, Sebastian Stein

School of Electronics and Computer Science
University of Southampton, UK
{j.williams, v.yazdanpahah, s.stein}@soton.ac.uk

ABSTRACT

In this paper, we introduce an audio-based framework for occupancy estimation, including a new public dataset, and evaluate occupancy in a 'cocktail party' scenario where the party is simulated by mixing audio to produce speech with overlapping talkers (1-10 people). To estimate the number of speakers in an audio clip, we explored five different types of speech signal features and trained several versions of our model using convolutional neural networks (CNNs). Further, we adapted the framework to be privacy-preserving by making random perturbations of audio frames in order to conceal speech content and speaker identity. We show that some of our privacy-preserving features perform better at occupancy estimation than original waveforms. We analyse privacy further using two adversarial tasks: speaker recognition and speech recognition. Our privacy-preserving models can estimate the number of speakers in the simulated cocktail party clips within 1-2 persons based on a mean-square error (MSE) of 0.9-1.6 and we achieve up to 34.9% classification accuracy while preserving speech content privacy. However, it is still possible for an attacker to identify individual speakers, which motivates further work in this area.

***Index Terms*—** occupancy, privacy, speaker identification, safe AI, speaker recognition, smart buildings

1. INTRODUCTION

One of the foremost research problems for smart buildings is occupancy estimation: counting the number of people in a room or building. Occupancy estimation algorithms interface directly with AI-based building management systems that aim to minimise energy use while ensuring occupant comfort and welfare [1]. As a type of sensor for occupancy estimation, audio-based solutions are under-explored even though they offer a cost-effective solution that can be interlinked with additional smart building audio AI services including audio scene understanding [2] or to enhance security using voice biometrics for accessing restricted areas [3]. Moreover, many buildings already contain the necessary hardware as people increasingly adopt audio-enabled devices in their homes such as smart voice assistants or smart TVs.

Early work on occupancy estimation from audio has treated this problem too coarsely to be used in real-world deployments and datasets were not publicly available. For example, when occupancy is treated as a binary problem (i.e., a room is either occupied or not), the binary occupancy information does not fully enhance AI building management systems [4]. Estimated counts can also vary greatly in magnitude (10 vs 200 people) [5] or can be limited to very small counts in specific scenarios as with 3 occupants inside of a small residence [6].

Furthermore, audio-based occupancy estimation offers more opportunities for privacy preserving approaches and is less computationally expensive than video [6]. Audio also overcomes issues of *line of sight* that arise from other types of sensors such as video, passive infrared (PID) and ultrasonic sensors [7]. To the best of our knowledge, we are the first to explore intentional privacy-preserving strategies for the problem of audio occupancy estimation for overlapping speakers. This scenario is applicable to multiple domains where people may speak simultaneously such as open-office layouts, conferences, museums, and cafes. Our work explicitly accounts for privacy of individuals and conversations by using strategies of audio degradation before training classifiers [8]. We adopt a technique from [9] called *shredding* to slice and re-arrange audio frames. We then combine that technique with *random reversal* in the time-domain to establish our baseline estimates to count the number of speakers. Our contributions in this paper are as follows:

1. Introduce a new dataset that simulates overlapping speakers at a 'cocktail party' or similar environment such as a cafe or conference.
2. Establish a set of baselines with and without privacy-preserving strategies for occupant estimation using several different types of audio features.
3. Probe the efficacy of privacy safeguards using occupancy estimation features on two adversarial tasks: automatic speaker verification (ASV) and automatic speech recognition (ASR).

2. BACKGROUND

There has been some limited previous work regarding the use of audio solutions for occupancy detection. Audio was used for occupancy detection in [6], however the dataset that was collected and evaluated in the paper is considered to be proprietary and cannot be shared for replication experiments or further research. Their work explored presence detection (as a binary classification problem) and head count (as a multi-class classification problem) on two types of rooms: living room and a single office. They extracted low-level signal features such as zero crossing rate, spectral variance and Mel-frequency cepstral coefficients (MFCCs) using the least absolute shrinkage and selection operator (LASSO). They report a high accuracy for presence detection of 99% and an accuracy of 70% for head count. While their study explored only a small number of occupants, and although they argue that audio is less intrusive than video in their work, they did not explicitly address privacy through experimentation.

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the Trustworthy Autonomous Systems Hub (EP/V00784X/1), a Turing AI Acceleration Fellowship on Citizen-Centric AI Systems (EP/V022067/1), and by the Southampton Low Carbon Comfort Centre.

Motivated by the needs of green energy and smart resource management in buildings, recent work has examined how various occupancy sensors can have a positive impact on resources. In [10], the authors discuss that audio is an under-studied sensor for occupancy estimation even though it offers a low start-up cost and is easy to retrofit to buildings. While audio does not suffer from line-of-sight issues, it can be hindered by environmental sounds that interfere with occupancy estimation algorithms. Their work explores sound cancellation to overcome this limitation. They mixed clean speech with additive Gaussian white noise. Their methodology assumes that speakers always take turns, speaking one by one. They examine scenarios where there are either 10, 20, or 40 speakers in the audio and used a speaker recognition technique based on Gaussian mixture models (GMMs) for occupancy estimation. They found that noise significantly degrades classification performance, but the performance improves when using speech enhancement. Our work in this paper makes slightly different assumptions than [10], namely that speakers do not take turns. We also examine a finer-grained problem of counting occupants from 1-10 speakers.

Very recent work from [5] explored the use of low-level signal descriptors (e.g., dominant frequency, loudness in dB, and reverberation time) for occupancy and activity detection in audio. They experimented with the DISCO dataset[1] [11], which contains 2000 images and corresponding 1 second of audio, but they used only the audio portion for experiments. The occupant counts in the DISCO dataset range from 0-700+ and were hand-labeled using the images as ground truth. They fit a regression model and measured mean absolute error (MAE), finding that their algorithm *XGBoost* achieved best performance of $MAE = 49.06$. For very high occupant counts, this may be an acceptable error depending on the use-case. Importantly, we do not use speaker recognition techniques in our occupancy estimation techniques as these are not aligned to our goals of privacy-preservation.

3. DATASET CREATION AND AUDIO FEATURES

In this section, we introduce the dataset that we created for simulating multiple 'cocktail party' scenarios. We further detail the audio features that were used in classification experiments and the privacy-preserving data degradation strategies.

3.1. Simulated Cocktail Party Dataset Using LibriTTS

We created a cocktail party dataset[2] by mixing audio from speakers in the LibriTTS [12] dataset so that they overlapped, simulating how multiple people may speak simultaneously[3]. LibriTTS is a subset of LibriSpeech [13]. We selected LibriTTS-Clean360 and LibriTTS-Clean100, and the development set because these had undergone pre-processing and cleaning before being released. All audio files were downsampled to 16 kHz and converted to 16-bit PCM wave format. The waveform amplitude was normalized to -26 dBov in advance using ITU-T G.191 sv56 [14]. We set aside 804 speakers for the training and development sets and 201 held-out speakers for the test set. Then we created different mixtures of overlapping speakers using SoX[4], by randomly selecting N speakers in the range of [1,10] and 1 utterance per speaker. Thus, samples with the fewest number of speakers contained only 1 speaker, while samples of the highest number of speakers contained 10 speakers. After mixing, the samples we re-normalized to -26 dBov and the file length trimmed according to the shortest utterance in the mixture so that the file duration always contains overlapping speakers without trailing silence. The duration of each mixed audio file in the dataset was between 2 and 5 seconds. This duration range was intentionally utilized to mimic how a smart building may use microphones to sporadically sample occupancy levels while minimizing privacy concerns that a building could be 'always listening'.

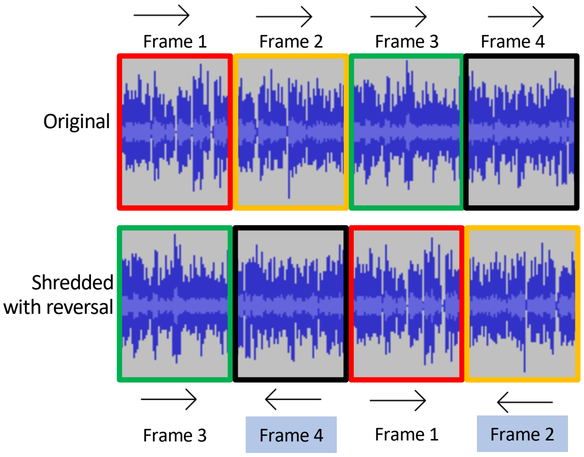

Fig. 1. Demonstration of shredding to re-order audio frames (128ms) and random reversal to reverse audio frames in the time domain.

3.2. Privacy from Data Degradation

As proposed by [8] and [9], *data degradation* is a type of privacy-preserving strategy that aims to limit how input features can be used in certain adversarial types of tasks while enforcing acceptable performance on a target task. In this work, our target task is occupancy estimation and the adversarial tasks are ASV and ASR which could violate privacy expectation. Therefore the data degradation strategies that we experiment with are meant to perform very well for occupancy estimation, and very poorly on ASV and ASR.

The first data degradation technique called *sound shredding* comes from [9], which applied it to detecting background environments (e.g., meeting, going for a walk, cafe, laboratory, etc). The technique was shown to help reduce human perception of speech content, speaker identity, and speaker gender. We are the first to apply shredding as a privacy-preserving mechanism for the problem of occupancy estimation. For shredding, we divided the raw waveform into 128ms frames and randomly shuffled all of the frames. We used 128ms frames because phones in speech at a normal speaking rate are on average ˜150ms duration [15], so this would potentially degrade the raw audio enough to disguise spoken content. We introduce a second data degradation technique called *random reversal*, which selects 50% of the shredded frames, and reverses them in the time domain. We included this technique aiming to enhance privacy for both speech content and speaker identity. Figure 1 demonstrates both shredding and reversal as data degradation techniques applied to original waveforms.

[1] https://github.com/qingzwang/AudioVisualCrowdCounting

[2] http://www.openslr.org/135

[3] https://rhoposit.github.io/icassp2023c

[4] http://sox.sourceforge.net

	Original		Shredded		Shredded + Reversal	
	%ACC ↑	MSE ↓	%ACC ↑	MSE ↓	%ACC ↑	MSE ↓
Waveform	**30.9±24.6**	**0.8±0.8**	**34.9±28.4**	**0.9±0.8**	**28.7±24.7**	**0.9±0.8**
MFCC	20.9±10.9	1.3±0.7	30.8±24.9	1.4±0.9	20.9±11.3	1.5±0.7
LFCC	27.5±28.0	1.8±1.1	28.0±27.1	1.8±1.0	28.3±26.4	1.5±0.8
Magspec	16.2±10.0	2.1±1.1	19.1±12.0	2.1±1.1	22.8±24.8	2.1±1.1
Melspec	16.5±14.6	2.6±1.8	19.0±10.4	1.7±1.0	23.4±24.7	1.7±1.0

Table 1. Average per-class accuracy (ACC) and mean-square error (MSE) on held-out test using original audio, shredded audio, and shredded audio combined with random reversal.

3.3. Features for Machine Learning Experiments

We utilized a variety of features extracted from the audio, including full waveforms. In particular, we used the following features: waveforms, magnitude spectrograms (*magspec*), Mel-frequency spectrograms (*melspec*), Mel-frequency cepstral coefficients (MFCCs), and linear-frequency cepstral coefficients (LFCCs). All features were extracted using libraries from TorchAudio [16]. When combining these feature types with the data degradation strategies, shredding and random reversal was applied to the waveform before feature extraction. For features involving full waveforms with privacy strategies, shredding and random reversal were performed after mixing speakers in the cocktail party scenarios.

4. EXPERIMENTS

We experimented with the various features and data-degradation strategies to predict the number of speakers in each audio file from our 'cocktail party' dataset. In this section, we introduce our machine learning experiments in terms of the architecture, hyper-parameters, and data training/testing partitions.

The architecture that we chose for experiments was a fully-convolutional CNN architecture inspired by [17] without any fully-connected layers. It was shown that this architecture is capable of analyzing environmental sounds and classifying speech directly from a waveform with minimal pre-processing. Fully-convolutional CNNs are well-suited to speech processing because waveforms are extremely high-dimensional inputs with information distributed throughout the higher dimensions.

We explored several different hyper-parameters using grid search: learning rate in the range of $\{0.01, 0.001, 0.0001\}$, weight decay (L2 normalization) $\{\text{None}, 0.01, 0.001, 0.0001\}$, the number 1D convolutional layers $\{1, 2, 3\}$, and batch sizes $\{128, 256\}$. Each convolutional layer also used a corresponding 1D max-pooling layer followed by a dropout layer (dropout=0.2). The CNNs were trained using early stopping with patience set to 5 epochs while monitoring validation accuracy.

We used portions of our simulated 'cocktail party' dataset to create training/validation sets and a held-out test set using 1000/200/200 instances for each of the 10 classes. This resulted in a dataset of 10k items for training, 2k for validation, and 2k for test. The training/validation partition included mixed audio from 804 speakers, while the test set included mixed audio from 201 held-out speakers. All of the speakers used in the held-out test set were unseen during training.

5. RESULTS

In Table 1, we report the results from classification experiments measured by accuracy (on the 10-class prediction problem) as well as mean-square error (MSE) between true and predicted ordinal classes (on a scale of 1-10) to further characterize performance. The results include experiments involving original audio and features without any privacy-preserving techniques, as well as two experiments wherein audio had undergone degradation before feature extraction and training. On this 10-class problem, we calculated accuracy for random guessing as $acc = 10\%$. For MSE, we generated random numbers in the range of [1,10] such that there were 200 exemplars for each of the 10 classes. We then calculated the MSE random baseline to be $mse = 20.5 \pm 9.8$.

In most experiments, the full waveform performed best for predicting the number of speakers present in our simulated cocktail party dataset, and better than the random baselines. The highest accuracy was 34.9% for MFCCs that were derived from waveforms that had undergone shredding. When measuring MSE, all three types of inputs (original, shredded, and shredded+reversal) indicate that the CNN can predict speaker counts within 1 person, which is good performance for our use-case of occupancy estimation.

Other features, such as magnitude spectrograms and Mel-frequency spectrograms did not achieve high classification accuracy or suitable MSE. Likewise, MFCCs and LFCCs did not perform especially well, except for the high classification accuracy for MFCCs extracted from shredded waveforms. This result was somewhat unexpected since MFCCs and LFCCs are features known to be useful in speaker recognition tasks [18]. It could be that occupancy estimation benefits more from high dimensional data or patterns from the time-domain, or that these other features are better-suited to feed-forward neural network architectures. We had expected both MFCC and LFCC to be especially relevant for our use case of privacy-preserving occupancy estimation for smart buildings that have been outfitted with microphones. Processing full waveforms may come with risks of privacy intrusion if data must be transferred away from the initial sensor (i.e., microphone) or digital signal processing (DSP) unit. However, the data degradation techniques of shredding and random reversal could be implemented at the sensor-level to mitigate privacy risks.

6. ANALYSIS

6.1. Speaker Recognition

The purpose of examining speaker recognition is to understand whether data degradation has preserved speaker-identifying characteristics in the audio. To evaluate this, we first extract utterance-level embeddings for each of the original speaker utterances from LibriTTS, for the 201 speakers and utterances in our held-out test set. Embeddings of *192-dim* are x-vectors extracted from each utterance using the pre-trained ECAPA-TDNN [19] system provided by SpeechBrain[5] [20] that was trained on VoxCeleb data [21]. We re-

[5] https://github.com/speechbrain/speechbrain

peat this again by extracting embeddings for each of the mixed audio files in the test set, for speaker counts of $\{1, 2, 3\}$. We then repeat the same for two privacy-preserving features: *shredded waveforms* and *shredded waveforms + random reversal*.

Next, we create trials to evaluate speaker recognition for each of the three features: full waveform, shredded, and shredded + random reversal. Embeddings from original audio for each speaker are treated as enrollment utterances and embeddings from mixed audio are treated as test utterances. We use 5 matched and 5 unmatched trials per speaker. For example, to evaluate speaker recognition in mixed audio that contains 2 speakers, we use 5 matched and 5 unmatched trials for each of the 2 speakers from the mixed file (20 trials total). Mixed audio with higher speaker counts therefore had a greater number of trials. We then repeat this for the shredded and shredded + reversal waveforms. The matched and unmatched trials were text-independent. Unmatched conditions utilized utterances from speakers that were not part of our train/valid/test split. Trials were evaluated using equal-error rate (EER) based on cosine similarity between embeddings. We expected that it is easier to identify unique speakers when the occupancy count is 1 or 2 speakers, and that it would be more difficult to identify speakers under conditions of data degradation. We experimented with using enrollment utterances from original LibriTTS audio as well as the degraded versions of enrollment utterances. Table 2 shows the resulting EER from these trials. Since we are interested in privacy, and the ability to *not* identify speakers, we are interested in higher EER values, especially for audio with a single speaker.

	Original Enrollment			Privacy Enrollment	
#Spk	Orig.	S	S+R	S	S+R
1	9.7	10.3	11.5	10.6	11.9
2	26.1	27.1	27.9	27.7	27.7
3	35.0	35.2	36.0	34.7	36.2

Table 2. Equal-error rate (EER↑) for each occupancy level (number of speakers in mixed audio) calculated from original waveform (Orig.), shredded waveform (S), and shredded waveform with 50% of frames reversed in the time domain (S+R).

The SpeechBrain x-vectors with cosine similarity scoring achieved 9.7% EER for single-speaker audio using the original full waveform for both enrollment and test utterances. This value is likely a consequence of the ASV system and the type of data that it was trained on. We intended for the degraded waveforms to perform very poorly on this task, however our attempts to conceal speaker identity did not improve with significant data degradation. Comparing the original enrollment, we can see that regardless of whether an adversary knows the privacy policy, they still may be able to identify speakers as if the audio was not degraded. These scores could potentially change if a different ASV system is tested or if a different scoring metric is used (such as PLDA) [22]. Our findings on this task suggest that more privacy-preserving mechanisms are needed to protect speaker identity for occupancy estimation problems.

6.2. Speech Recognition

We also investigated how well speech can be obtained from the original and degraded waveforms using ASR. We used the IBM Watson Speech-to-text API[6] to obtain transcripts from the audio. As with speaker recognition, we explored speaker counts of $\{1, 2, 3\}$, as it is very difficult to recognize speech when speaker counts are very high in the mixed audio. These lower speaker counts apply to use-cases where there may be 1 or 2 people in a room (e.g., on a phone call or chatting). After obtaining the ASR transcripts from the API, we measured speech recognition performance using match error rate (MER) [23, 24] calculated with the Python jiwer library[7]. Unlike word error rate (WER), the MER scoring measures the proportion of word substitutions, insertions, and deletions compared to the total number of words, and is interpreted as the probability that two strings do not match. MER values of $mer = 0$ indicate no errors while values of $mer = 1$ indicate that no words match. Thus, MER is dependent on edit cost between two strings.

In this analysis, we expect that the MER scores will become worse for higher speaker counts using the original waveforms. We also tested whether scores would be higher using the privacy-preserving techniques. The results for MER scoring are presented in Table 3, showing scores for original, shredded, and shredded + reversed waveforms. Both of our data-degradation strategies had a significant impact on ASR performance, rendering words unintelligible compared to the original waveform.

#Spk	Original	Shredded	Shredded+ Reversal
1	0.13	0.94	0.95
2	0.88	0.97	0.98
3	0.97	0.98	0.98

Table 3. Match error rate (MER↑) for each speaker counts calculated from original, shredded, and shredded + reversal waveforms.

7. DISCUSSION AND FUTURE WORK

We have presented a privacy-preserving framework for estimating people counts in audio where people are talking at the same time, as in a ‘cocktail party’ scenario. We applied and expanded data-degradation techniques to this type of problem and are the first to do so. We also created a new dataset of mixed audio which will be publicly released for future research. We achieved estimated speaker counts within 1-2 persons while preserving privacy based on speech recognition. However, our adversarial speaker recognition tasks indicate that more privacy techniques are needed to protect identity.

This study presents foundational work and therefore is not without limitations.We did not experiment with room simulation [25]. In addition, there are other audio features that may be interesting or relevant, such as features and perturbations for privacy in the frequency domain. In our experiments, we did not include a zero-occupancy condition but plan to evaluate the technique on a dataset such as DESED[8] that contains a variety of acoustic scenes with and without speech. This would allow us to evaluate whether the model is learning information based on human speech or simply the energy in the signal, and this is an important distinction to avoid false positives due to loud non-human sounds. Our approach does not account for maintaining state over a period of time, for example people entering or leaving a room, nor does it account for multiple talkers who take turns, for example in a meeting. As previously existing techniques aim to identify speakers explicitly, the role of speaker recognition is paramount to privacy for occupancy estimation as well as all audio-based technologies to be installed into smart buildings.

[6] https://www.ibm.com/uk-en/cloud/watson-speech-to-text

[7] https://pypi.org/project/jiwer/

[8] https://github.com/turpaultn/DESED/

8. REFERENCES

[1] Nadine von Frankenberg, Vivian Loftness, and Bernd Bruegge, "I want it that way: Thermal desirability in shared spaces," 2021, BuildSys '21, p. 204–207.

[2] Ziming Li, Shannon Connell, Wendy Dannels, and Roshan Peiris, "SoundVizVR: Sound Indicators for Accessible Sounds in Virtual Reality for Deaf or Hard-of-Hearing Users," in *Conference on Computers and Accessibility (ASSETS'22)*, 2022.

[3] Choon Beng Tan, Mohd Hanafi Ahmad Hijazi, Norazlina Khamis, Zuraini Zainol, Frans Coenen, Abdullah Gani, et al., "A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction," *Multimedia Tools and Applications*, vol. 80, no. 21, pp. 32725–32762, 2021.

[4] Bingqing Chen, Zicheng Cai, and Mario Bergés, "Gnu-RL: A precocial reinforcement learning solution for building HVAC control using a differentiable MPC policy," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, 2019, pp. 316–325.

[5] Gabriela Santiago, Marvin Jiménez, Jose Aguilar, and Edwin Montoya, "Audio feature engineering for occupancy and activity estimation in smart buildings," *Electronics*, vol. 10, no. 21, pp. 2599, 2021.

[6] Shabnam Ghaffarzadegan, Attila Reiss, Mirko Ruhs, Robert Duerichen, and Zhe Feng, "Occupancy detection in commercial and residential environments using audio signal," in *Proc. of Interspeech*, 2017, pp. 3802–3806.

[7] Zhenghua Chen, Chaoyang Jiang, and Lihua Xie, "Building occupancy estimation and detection: A review," *Energy and Buildings*, vol. 169, pp. 260–270, 2018.

[8] Jennifer Williams, Vahid Yazdanpanah, and Sebastian Stein, "Safe audio AI services in smart buildings," *9th ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys 2022)*, 2022.

[9] Sumeet Kumar, Le T Nguyen, Ming Zeng, Kate Liu, and Joy Zhang, "Sound shredding: Privacy preserved audio sensing," in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, 2015, pp. 135–140.

[10] Qian Huang, "Occupancy-driven energy-efficient buildings using audio processing with background sound cancellation," *Buildings*, vol. 8, no. 6, pp. 78, 2018.

[11] Hu Di, LiChao Mou, Qingzhong Wang, Junyu Gao, Yuansheng Hua, Dejing Dou, and Xiao Xiang Zhu, "Ambient sound helps: Audiovisual crowd counting in extreme conditions," in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1–4.

[12] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," *Proc. Interspeech*, pp. 1526–1530, 2019.

[13] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[14] International Telecommunication Union, ," Recommendation G.191: Software Tools and Audio Coding Standardization, Nov 11 2005.

[15] Hisao Kuwabara, "Acoustic properties of phonemes in continuous speech for different speaking rate," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, 1996, vol. 4, pp. 2435–2438.

[16] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, and et al., "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.

[17] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das, "Very deep convolutional neural networks for raw waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 421–425.

[18] Xinhui Zhou, Daniel Garcia-Romero, Ramani Duraiswami, Carol Espy-Wilson, and Shihab Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 2011, pp. 559–564.

[19] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[20] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, and et al., "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[21] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, Francisco Lacerda, Ed. 2017, pp. 2616–2620, ISCA.

[22] Qiongqiong Wang, Kong Aik Lee, and Tianchi Liu, "Scoring of Large-Margin Embeddings for Speaker Verification: Cosine or PLDA?," in *Proc. Interspeech 2022*, 2022, pp. 600–604.

[23] Andrew Cameron Morris, Viktoria Maier, and Phil Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," in *Eighth International Conference on Spoken Language Processing*, 2004.

[24] Foteini Filippidou and Lefteris Moussiades, "A benchmarking of IBM, Google and Wit automatic speech recognition systems," in *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 73–82.

[25] David Diaz-Guerra, Antonio Miguel, and Jose R Beltran, "gpurir: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.