INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Configurable Privacy-Preserving Automatic Speech Recognition

Ranya Aloufi, Hamed Haddadi, David Boyle

Imperial College London, UK
{ra6018,h.haddadi,david.boyle}@imperial.ac.uk

Abstract

Voice assistive technologies have given rise to far-reaching privacy and security concerns. In this paper we investigate whether modular automatic speech recognition (ASR) can improve privacy in voice assistive systems by combining independently trained separation, recognition, and discretization modules to design configurable privacy-preserving ASR systems. We evaluate privacy concerns and the effects of applying various state-of-the-art techniques at each stage of the system, and report results using task-specific metrics (*i.e.*, WER, ABX, and accuracy). We show that overlapping speech inputs to ASR systems present further privacy concerns, and how these may be mitigated using speech separation and optimization techniques. Our discretization module is shown to minimize paralinguistics privacy leakage from ASR acoustic models to levels commensurate with random guessing. We show that voice privacy can be *configurable*, and argue this presents new opportunities for privacy-preserving applications incorporating ASR.

Index Terms: speech recognition, privacy, human-computer interaction, computational paralinguistics

1. Introduction

Voice assistive technologies facilitate seamless interactions between consumers and online services using Automatic Speech Recognition (ASR) technology. Deep learning has been a driving force in research and practice across speech application domains, yielding rapid advancements in ASR. Although deep models have comparable performance with more conventional approaches like hidden Markov model (HMM)-based ASR, considerable privacy vulnerabilities arise due to training models on real voice data that contain a significant amount of sensitive information [1].

In this paper, we propose a modular system for privacy-preserving processing of voice data that combines independently trained separation, ASR, and discretization modules (Fig. 1). We study the effect of different state-of-the-art models at each stage of the pipeline and report results using task-specific metrics including word error rate (WER), zero-shot linguistics metrics [2] and downstream classification accuracy. The architecture must be modular to enable different privacy settings, thus we investigate the use of speech separation and speech discretization in an attempt to enhance privacy protection. Because the modules interact only at the I/O level, they can be implemented using different tools and independently optimized. Specifically, an obtained voice signal can be processed either by the Separation module to obtain a filtered signal belonging to a speaker of interest, or it may progress directly to ASR. The output of the separation module is fed to the ASR module. ASR systems consist of an encoder and a decoder to extract textual content from the voice signal. The output from the ASR module can be transferred to the Discretization module, which uses the output from the encoder of the ASR to extract the phonemic units (*i.e.* discrete, low bitrate, or pseudo-text units). Learning these discrete units can be used to extract a transcription directly by training a language model over it, or to obtain a new voice signal using a decoder trained on a system speaker (*i.e.* speech re-synthesis) [3].

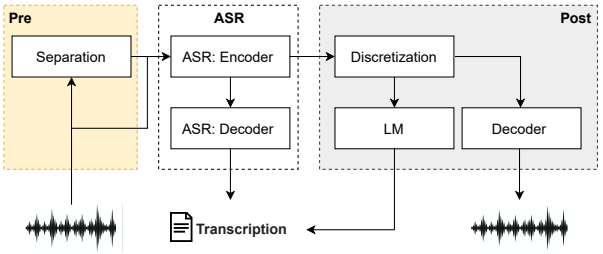

Figure 1: *Configurable privacy-preserving ASR architecture*

The **main contribution** of this paper is a configurable framework that enhances privacy in ASR systems based on speech separation and a discrete modular approach[1]. A new benchmark is defined against which to compare possible privacy settings. We also introduce phoneticsMix, a new speech dataset derived from Libri-light and WHAM [4] noise to evaluate ASR (*i.e.* ABX) in mixed utterance and zero-resource settings. We experimentally evaluate the proposed approach and perform a systematic performance analysis for task-specific metrics. We show that Separation improves WER by $\sim$16% in multi-speaker settings, and Discretization reduces classification accuracy in inferring sensitive attributes to 11%-51% (*i.e.* over multi/binary attributes).

2. Voice Privacy

2.1. Privacy-preserving Voice Analytics

Privacy enhancing technology for voice assistive systems is an emerging and important area of research. Aloufi *et al.* describe scenarios whereby attackers can infer a significant amount of private information by observing the output of state-of-art deep acoustic models for speech processing tasks [5]. Nautsch *et al.*, in [6], demonstrate the importance of proposing and developing new privacy-preserving technologies for protecting speakers and speech characterization in voice signals.

The recent VoicePrivacy initiative [7] promotes the development of anonymization methods that aim to suppress personally identifiable information in speech (*i.e.* speaker identity) while leaving linguistic content intact. Most of the proposed works focus on protecting/anonymizing speaker identity using voice conversion (VC) mechanisms [8, 9, 10]. These VC methods, however, aim to protect the speaker identity against different linkage attacks limited by the attacker's knowledge [11]. It

[1]Models and code will be publicly released on acceptance https://github.com/RanyaJumah/EDGY

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-1783

has been found that when the attacker has complete knowledge of the VC scheme and target speaker mapping, none of the existing VC methods can protect speaker identity. Beyond speaker identity, various works propose to protect speaker gender [12] and emotion [13], wherein an edge-based system is proposed to filter affect patterns from a user's voice before sharing it with cloud services for further analysis.

Another direction is protecting users' privacy by ensuring that sensitive data is not unnecessarily transmitted to service providers. This may be done by optimizing the neural network architecture using quantization and pruning techniques to enable on-device processing. He *et al.*, for example, propose an end-to-end speech recognizer for on-device applications such as voice commands and voice search that runs comfortably in real time on a Google Pixel phone [14]. Similarly, Aloufi *et al.* analyze optimization techniques including filter-pruning, weight-pruning, and quantization to learn privacy-preserving representations from raw data in near real-time with minimal model performance loss [15].

2.2. Potential Threat

Voice-controlled technologies that allow users to speak to interact with their devices are based on accurate speech recognition to ensure appropriate responsiveness. In many real-world usage scenarios, voice inputs may often consist of overlapping speech. This raises privacy concerns owing to the capture and processing of audio that may involve two or more people, and without their explicit consent. This violates GDPR provisions, for instance. Furthermore, the deep acoustics models used to analyze these recordings may encode more information than needed for the task of interest (*i.e.* ASR), such as profiles of users' demographic categories, personal preferences, emotional states, etc., and may therefore significantly compromise their privacy.

Protecting privacy requires more than hiding speaker information or running on-device ASR. We consider that privacy is subjective, with varying attitudes between users, which may even depend on the services (and/or service providers) with which these systems communicate. We therefore advocate the principle of *configurable privacy*. Unlike other approaches, we seek to protect the privacy of multiple user attributes for various scenarios that depend on voice input or speech analysis, *i.e.* sanitizing the speech signal of attributes a user may not wish to share, but without adversely affecting functionality. We also emphasize the importance of enabling different privacy settings for optimizing the privacy-utility trade-off and promoting transparent privacy management practises.

3. Configurable Privacy ASR

3.1. Speech Separation & Discretization

Personal voice activity detection (VAD) [16] systems detect the voice activity of a target speaker at the frame level, only triggering for the target user. This helps reduce computational cost and energy consumption, particularly in scenarios where a keyword detector is not preferable. VoiceFilter-Lite [17] is a single-channel source separation model that runs on-device to preserve only the speech signals from a target user as part of a streaming speech recognition system. Similarly, Xue *et al.* in [18] propose a method called speaker tracing buffer, which can track speaker information consistently across the chunk by extending a self-attention mechanism to maintain the speaker permutation information determined in previous chunks. Target-speaker voice activity detection and E2E speaker-attributed ASR can thus play a significant role in improving E2E ASR performance on overlapped speech, as well mitigating privacy concerns.

Discovering discrete units may be useful for phonetic learning for low resourced languages or for languages with no textual resources, which cannot be addressed in the supervised setting, making speech technology more inclusive. One motivation behind discretization is that it can capture high-level semantic content from the speech signal, *e.g.* phoneme, due to the discrete nature of phonetic units [2]. Such representations could be used in downstream speech applications requiring symbolic or sparse input, *e.g.* for faster retrieval in speech search systems [19] or to build a synthesizer in a target speaker's voice using the discovered units [20].

3.2. Design Overview

Our pipeline consists of three basic components: Separation (SEP), ASR, and Discretization (DIS), as shown in Figure 1.

3.2.1. Data & Models

We evaluate the proposed approach using a number of real-world datasets that were recorded for various speech processing purposes. We use the dev clean and other sets of Libri-light [21] to evaluate the quality of discovering the discrete units from the raw recordings. We use LibriCSS [22] and phoneticsMix to investigate the influence of overlapping speech in the raw recordings on E2E ASR systems and the separation role to reduce it. Inspired by LibriMix [23], we introduce phoneticsMix to investigate the effect of overlap speech on learning discrete units. The phoneticsMix dataset is derived from Libri-light, which was introduced mainly for training speech recognition systems under limited or no supervision (*i.e.* providing phonetic sequences as ground-truth), which we mixed with WHAM [4] noise to investigate speech separation for zero resources ASR that discovers its own units from raw speech. We use additional sets for various paralinguistics purposes, including speaker recognition (VoxCeleb [24]), accents and gender (Common Voice [25]), and emotion recognition (CREMA-D [26] and SAVEE [27]) to estimate the privacy leakage potentially caused by deep acoustics models of ASR systems.

We conduct our experiments using various state-of-the-art (SOTA, Table 3) models per module in the proposed framework. We first use SepFormer [28] as a speech separation model which is an RNN-free architecture that employs a masking network composed of transformers only. We then implement our ASR experiments on CRDNN with connectionist temporal classification (CTC)/Attention-based and a Transformer-based E2E ASR model trained on Librispeech within SpeechBrain [29]. To enable speech discretization, we follow Zero Resource 2021 baseline [2] by concatenating self-supervised representation learning encoder with k-means clustering to convert continuous representations into discrete representations. We use the pre-trained contrast predictive coding (CPC) [2] and wav2vec 2.0 [30] to encode the raw recordings into continuous representations and then train k-means on top of these representations.

3.2.2. Speech Separation

Q1: *a.) How does overlapping speech affect users' privacy? b.) Can speech separation help enhance ASR performance and accurately identify the speaker of interest to promote further privacy-preserving applications?*

Most prior studies concerning voice signal privacy have

Table 1: *Overall performance of zero resource ASR on clean and other (O) sets on ABX zero-shot metric (M:Mix, lower ABX is better)*

	ABX within				ABX across			
Model	Dev-clean	Dev-O	DevM-clean	DevM-O	Dev-clean	Dev-O	DevM-clean	DevM-O
CPC+DIS	10.26	14.24	14.11	16.24	14.17	21.26	20.37	24.06
SEP+CPC+DIS	06.38	08.22	13.19	14.57	10.22	14.84	18.60	21.11
wav2vec2+DIS	16.47	24.30	28.41	30.78	17.41	27.43	20.86	29.28
SEP+wav2vec2+DIS	14.09	20.28	20.21	27.46	19.51	27.26	20.71	29.46

Table 2: *Overall performance of E2E ASR models on LibriCSS using WER metric (0S and 0L: 0% overlap with short and long inter-utterance silence respectively)*

	Overlap ratio in %					
Method	0L	0S	10	20	30	40
CRDNN	13.9	18.8	25.9	31.4	38.3	43.9
SEP+CRDNN	11.4	15.3	21.7	27.3	34.4	40.4
TRANS	14.4	16.4	23.5	33.0	35.3	45.2
SEP + TRANS	8.2	14.5	16.5	19.2	26.6	29.4

been constrained to mixtures of clean, near anechoic speech, not representative of many real-world scenarios. Overlapping speech influences a significant part of conversational interactions, and this may pose challenges for many speech technologies including E2E ASR and speaker verification. This is because they usually assume one or zero speakers to be active at the same time. Speech separation might therefore provide a solution for this overlapping speech problem. It might also lead to robust performance for voice assistive systems and a potential solution to improve privacy (*i.e.* by excluding the non-target speaker information). We are interested in a scenario where multiple speakers may be present during raw recording. We first evaluate the performance of state-of-the-art ASR models in transcribing mixed utterances. We assess the influence of mixed utterances in learning speech representations from the raw signals. We then apply a single-channel source speech separation model to extract the speech signals from a target user only (*i.e.* use only the first channel data (monaural audio) for all our experiments), assuming this single channel belongs to that user and estimate the effect of this module in enhancing the performance in waveform or representations levels.

Metric. We report the word error rate (WER) on LibriCSS [22] evaluation set to estimate the performance of the E2E ASR models (*i.e.* using waveform). For the representations level, we run the ABX metric [2] on the Libri-light dev clean and other sets to estimate the discriminability between phonemes (lower is better). ABX is calculated by computing the distance between the representations associated with three acoustic tokens (*a*, *b*, and *x*), two of which belong to the same category *A* (*a* and *x*) and one which belongs to a different category *B* (*b*). Thus, the score is the estimated probability that *a* and *x* are closer to one another than *a* and *b*. We compute it within speaker (in which case, all stimuli *a*, *b*, and *x* are uttered by the same speaker) and across speakers (*a* and *b* are from the same speaker and *x* from a different speaker).

Results. We conduct two types of experiments to measure the performance of the learning task, E2E ASR and zero-resource ASR, using clean and mixed utterances to simulate overlapped interactions in realistic voice assistive deployment environments.

No-SEP. No separation mechanism is applied. As shown in Table 1, the overlapped speech critically degraded the ASR performance in both settings (*i.e.* E2E and zero-resource) about 8%-40% WER and 6%-8% for E2E and zero-resource ASR, respectively. While we test it over mixed utterances (*i.e.* multi-speaker) only, this decrease may be worse with a noisier and more overlapped percentage.

With-SEP. Even a small amount of overlap (10%) critically degraded the ASR performance, however, the single-channel separation model might help improve its performance. The result is much better than the "No-SEP" scenario, showing that the single-channel separation model can improve the WER accuracy by 2%-16% in E2E ASR and about 2% ABX score in zero-resource ASR (see Table 1). The latter, however, needs more investigation to understand the effect of the mixed environments on the phonetic level, and more separation mechanisms may be needed for further improvement since zero-resource ASR developed under an unsupervised manner.

Once we tune this separation stage, the output can be connected to various applications, which could be speech recognition or speaker verification.

3.2.3. Speech Discretization

Q2: *a.) Can deep acoustic models of E2E ASR reveal sensitive attributes about users? b.) Can discrete units (e.g., phonemes) help to limit the amount of unnecessarily sensitive information learned by these models?*

Discrete units (*e.g.* phonemes) highlight linguistically relevant representations of the speech signal in a highly compact format [2] while being invariant to speaker-specific and background noise details. These additional acoustics information might lead to accurately inferring users' sensitive and private attributes (*e.g.* their gender, emotion, or health status). For example, an attacker (*e.g.* a 'curious' service provider) may use this information encoded by speech recognition or speaker verification system to learn further sensitive attributes from user input with high inferring accuracy ranging from 40% to 99.4%, significantly better than guessing at random [5].

One motivation to apply discretization is that these discrete units can capture high-level semantic content from the speech signal, *e.g.* phoneme due to the discrete nature of phonetic units. We investigate the effectiveness of *learning discrete units* in preserving the sensitive attributes in the encoded speech data. We extract an embedding using the encoder model of E2E ASR and self-supervised encoders (*i.e.* CPC and wav2vec2), and then discretize the output of these encoders using k-means. We use k-means as method to convert continuous frame representations into discrete representations using centroids of clusters obtained by training k-means on LibriSpeech. Each audio file will thus be discretized to a sequence of discrete units corresponding to the assigned clusters.

Metric. We estimate the level of the paralinguistics information preserved in encoded linguistics representations generated through ASR by measuring the classification accuracy of various paralinguistics tasks (*i.e.* speaker recognition, emotion recognition, accent identification, and gender recognition).

Table 3: *Performance for paralinguistic tasks (*Results incorporate a subset of Common Voice for accent/gender recognition)*

Task	Speaker Recognition	Emotion Recognition		Accent Identification	Gender Recognition
Model/Dataset	VoxCeleb1	CREMA-D	SAVEE	Common Voice*	Common Voice*
SOTA	80.5	74.0	68.5	-	-
CRDNN	39.07	52.23	66.40	36.54	52.82
TRANS	47.04	59.31	71.53	38.56	57.51
CPC	45.11	62.02	57.91	71.69	82.65
wav2vec2	37.04	51.49	51.52	40.31	52.5
CRDNN+DIS	19.03	20.31	16.91	19.93	43.82
TRANS+DIS	20.14	23.35	20.83	20.68	44.21
CPC+DIS	17.04	29.19	19.33	23.72	44.51
wav2vec2+DIS	14.54	25.88	18.15	15.06	41.32

Specifically, we test the accuracy in predicting these attributes over random guessing by training shallow models using Scikit-Learn on top of these representations.

Results. We first evaluate the quality of the linguistics representations by aiming for high performance in the linguistics task (Section 3.2.2), while achieving low performance on paralinguistics tasks (*i.e.* increase in inference accuracy over random guessing).

No-DIS. No discretization mechanism is applied. The accuracy of paralinguistics attributes in Table 3 (*i.e.* upper rows) shows that the possibility to infer speaker-sensitive information ranging from 36.5%-82.7% (*i.e.* for binary (gender) and multi-class (speaker, emotion, and accents) attributes), namely that a predictor of speaker characteristics can be learned from deep acoustics models used by ASR systems.

With-DIS. As shown in Table 3, the accuracy of the paralinguistics classification models dropped by 11%-51%, which is close to or less than for random guessing. For example, that the sensitive attribute in question is the speaker's 'emotion', we have seven labeled categories in the available datasets (CREMA-D and SAVEE). The random guess rate for success is therefore around 14%. If we assume that the sensitive attribute is 'gender' (*e.g.* binary male or female), the random guess rate will be 50%. Thus, such discretization may help to improve obtaining good language representations while reducing irrelevant paralinguistics information that may pose a threat to the user privacy.

Once we tune this discretization stage, the output can be connected to different downstream purposes, which could be direct transcription or speech resynthesis [3].

4. Discussion

Most recently proposed solutions for privacy preservation fall between anonymization and cryptography [7]. Although these solutions were effective in anatomizing the speaker identity, they may prevent speaker authentication regarding voice assistive systems activation. These solutions may be useful from singular perspectives and for achieving a specific goals, like the identity of the speaker, but might fail to sufficiently address configurable privacy. Therefore, we highlight the need for a 'configurable privacy' framework that can process the speech data in a privacy-preserving, task-related, and modular manner. First, we consider concern about overlapping speech on privacy, assuming the real-time deployment of voice assistive includes noisy and multi-speaker environments. Such environments could significantly affect the performance of ASR technology used by these assistive systems, leading to misactivations [31] or even open a door for further security attacks (*e.g.* adversarial examples [32]) to fool their prediction. We therefore show one potential mitigation by highlighting the importance of speech separation as a pre-step for voice assistive systems. As shown in Table 2, our experiments over mixed and noisy recordings show that these variations lead to a notable impact on E2E ASR, exhibiting $\sim$40% drop in WER, whereas using speech separation can help to improve the performance by 16%. In future work, we plan to extend the pre-step modules to include speaker diarization or even target speaker voice detection as a guide to the separation process and examine how that may improve privacy.

We also focus on the overlearning problem caused by deep acoustics models used for E2E ASR and speaker verification, revealing additional/sensitive information about the speaker from the raw recordings. The ability to discover the discrete units from raw signals and train language models (LMs) would go a long way towards making language technologies more inclusive, and encoding more acoustic information (*e.g.* speaker-specific) can be detrimental in achieving this purpose [2]. As shown in Table 3, the inference accuracy significantly decreased after implementing discretization by about 11%-51%. Learning discrete units can also help to protect these sensitive attributes. In addition, transferring these units might be useful for compressing and transmitting voice signals [33]. Discretization, therefore, might improve voice communication in terms of privacy and processing efficiency, especially for languages with limited resources.

5. Conclusions

We presented a modular framework that includes three basic components: Separation, ASR, and Discretization, to reconcile speech processing technologies and privacy restrictions for voice assistive systems. We studied the performance of E2E ASR under variation in recording equipment (e.g. noisy conditions) and multi-speaker conversational scenarios and showed speech separation may facilitate privacy-preserving applications. We then investigated speech discretization to minimize the privacy leakages against inferences of users' sensitive or private attributes. Our experiments indicate that using speech separation can help to improve performance by 16% and that discretization can effectively mitigate the overlearning issue caused by deep acoustics models. In the next steps of our work, we intend to extend our framework by consolidating recent developments with neural methods on speech processing technologies and progress towards more configurable, efficient, and privacy-preserving voice assistive technologies.

6. References

[1] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," *arXiv preprint arXiv:2001.00378*, 2020.

[2] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," *arXiv preprint arXiv:2011.11588*, 2020.

[3] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed, and E. Dupoux, "Generative spoken language modeling from raw audio," *arXiv preprint arXiv:2102.01192*, 2021.

[4] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," in *Proc. Interspeech*, Sep. 2019.

[5] R. Aloufi, H. Haddadi, and D. Boyle, "Privacy-preserving voice analysis via disentangled representations," in *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, ser. CCSW'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–14.

[6] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez-Barrero, D. Petrovska-Delacrétaz, G. Chollet, N. W. D. Evans, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Comput. Speech Lang.*, vol. 58, pp. 441–480, 2019.

[7] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé *et al.*, "Introducing the voiceprivacy initiative," *arXiv preprint arXiv:2005.01387*, 2020.

[8] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. Association for Computing Machinery, 2018, p. 82–94. [Online]. Available: https://doi.org/10.1145/3274783.3274855

[9] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," *arXiv preprint arXiv:2005.08601*, 2020.

[10] S. Ahmed, A. R. Chowdhury, K. Fawaz, and P. Ramanathan, "Preech: A system for privacy-preserving speech transcription," in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020, pp. 2703–2720.

[11] B. M. Lal Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2802–2806.

[12] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," *arXiv preprint arXiv:1910.13212*, 2019.

[13] R. Aloufi, H. Haddadi, and D. Boyle, "Emotion filtering at the edge," in *Proceedings of the 1st Workshop on Machine Learning on Edge in Sensor Systems*. Association for Computing Machinery, 2019.

[14] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.

[15] R. Aloufi, H. Haddadi, and D. Boyle, "Paralinguistic privacy protection at the edge," *arXiv preprint arXiv:2011.02930*, 2020.

[16] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I.-L. Moreno, "Personal vad: Speaker-conditioned voice activity detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 433–439.

[17] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika, and A. Gruenstein, "Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition," 2020.

[18] Y. Xue, S. Horiguchi, Y. Fujita, S. Watanabe, and K. Nagamatsu, "Online end-to-end neural diarization with speaker-tracing buffer," *arXiv preprint arXiv:2006.02616*, 2020.

[19] P. Peng, H. Kamper, and K. Livescu, "A correspondence variational autoencoder for unsupervised acoustic word embeddings," *arXiv preprint arXiv:2012.02221*, 2020.

[20] K. Kobayashi, W.-C. Huang, Y.-C. Wu, P. L. Tobing, T. Hayashi, and T. Toda, "crank: An open-source software for nonparallel voice conversion based on vector-quantized variational autoencoder," *arXiv preprint arXiv:2103.02858*, 2021.

[21] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673.

[22] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: dataset and analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7284–7288.

[23] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[24] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Interspeech 2017*, Aug 2017. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-950

[25] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[26] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[27] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[28] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," 2021.

[29] M. Ravanelli, T. Parcollet, A. Rouhe, P. Plantinga, E. Rastorgueva, L. Lugosch, N. Dawalatabad, C. Ju-Chieh, A. Heba, F. Grondin, W. Aris, C.-F. Liao, S. Cornell, S.-L. Yeh, H. Na, Y. Gao, S.-W. Fu, C. Subakan, R. De Mori, and Y. Bengio, "Speechbrain," https://github.com/speechbrain/speechbrain, 2021.

[30] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[31] D. J. Dubois, R. Kolcun, A. M. Mandalari, M. T. Paracha, D. Choffnes, and H. Haddadi, "When speakers are all ears: Characterizing misactivations of iot smart speakers," in *Proceedings of the 20th Privacy Enhancing Technologies Symposium (PETS 2020)*, 2020.

[32] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, and P. Traynor, "Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems," *arXiv e-prints*, pp. arXiv–2007, 2020.

[33] W. B. Kleijn, A. Storus, M. Chinen, T. Denton, F. S. C. Lim, A. Luebs, J. Skoglund, and H. Yeh, "Generative speech coding with predictive variance regularization," *arXiv preprint arXiv:2102.09660*, 2021.