INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

On-Device Speaker Anonymization of Acoustic Embeddings for ASR based on Flexible Location Gradient Reversal Layer

Md Asif Jalal[1], *Pablo Peso Parada*[1], *Jisi Zhang*[1], *Karthikeyan Saravanan*[1], *Mete Ozay*[1], *Myoungji Han*[2], *Jung In Lee*[2], *Seokyeong Jung*[2]

[1]Samsung Research, UK
[2]AI R&D Group, Samsung Electronics, Suwon, South Korea

mdasif.jalal@samsung.com

Abstract

Smart devices serviced by large-scale AI models necessitates user data transfer to the cloud for inference. For speech applications, this means transferring private user information, e.g., speaker identity. Our paper proposes a privacy-enhancing framework that targets speaker identity anonymization while preserving speech recognition accuracy for our downstream task - Automatic Speech Recognition (ASR). The proposed framework attaches flexible *gradient reversal based speaker adversarial layers* to target layers within an ASR model, where speaker adversarial training anonymizes acoustic embeddings generated by the targeted layers to remove speaker identity. We propose on-device deployment by execution of initial layers of the ASR model, and transmitting anonymized embeddings to the cloud, where the rest of the model is executed while preserving privacy. Experimental results show that our method efficiently reduces speaker recognition relative accuracy by 33%, and improves ASR performance by achieving 6.2% relative Word Error Rate (WER) reduction.

Index Terms: speech privacy, embedding privacy, embedding to audio synthesis, speech recognition.

1. Introduction

The increasing prevalence of voice driven human-computer interaction services in appliances has raised concern with regard to voice privacy and personal information protection. These 'smart' devices, ranging from cars to small watches, collect speech utterances and acoustic events for various downstream tasks or for training and evaluation in distributed settings [1]. Speech utterances hold user information such as speaker identity, gender etc. Privacy preservation is of critical importance to protect reliability in private data sharing.

Various privacy preservation methods for speech have been proposed in the literature. One solution is to manipulate speaker identity related features through feature perturbation [2], voice normalisation [3, 4], utterance slicing techniques [5], and differential pitch anonymization [6]. State-of-the-art methods employ neural based speech synthesizer or voice converter to generate speech where the speaker identity information has been removed [7, 8]. However, these methods require employment of additional synthesis modules and are computationally expensive, which is unrealistic for on-device scenarios.

An alternative approach for speaker anonymization is to learn speech representations invariant to speaker conditions. Domain adversarial training trains a model to learn domain agnostic representations [9]. Speaker based domain adversarial training has been effective for anonymizing latent representations of ASR models (i.e., acoustic embeddings) [10, 11]. However, it was observed that speaker invariant representations resulted in a reduction of ASR performance [10]. Orthogonal to this, recent work by [12] discuss a method where adding speaker-labels and adaptive gradient scaling to domain adversarial training improves ASR performance. However, they do not target or discuss privacy.

In this paper, we propose a flexible gradient reversal based speaker anonymization framework, which learns speaker anonymous acoustic embeddings within an ASR model while preserving its accuracy/performance (as depicted in Stage 1 in Figure 1). The initial layers of ASR models learn generic acoustic and prosody features, and the last layers learn more task-dependant semantic and syntax level features [13–15]. The research focuses on embeddings at the initial layers of ASR models. Furthermore, we introduce an acoustic embedding-to-waveform synthesis model to synthesise the corresponding audio waveform of the acoustic embedding for better understanding and interpretation (as shown in Stage 2 in Figure 1).

The main contributions of this paper are as follows:

1. We propose a method to use single gradient reversal at flexible layers of an ASR model to effectively mitigate speaker information from the representations generated by initial layers of the model without increasing its WER. In the analyses, we observed that speaker identification accuracy was reduced by 22% at layer 3 (CE3), 7.3% at layer 5 (CE5), and 6% at layer 7 (CE7) compared to the original speech waveform (Table 2). Performance of the models trained with these representations was improved by 8.6% WER on average. The proposed method does not require computationally expensive voice-conversion/speech-synthesis models for anonymization and operates on ASR embeddings.
2. Our results show that while having improved ASR performance, the speaker adversarial training has anonymized acoustic embeddings with gradient scaling. A detailed analysis of the effects of gradient scaling, domain loss scaling and model layer hierarchies are presented with performance of models and their convergence properties. Furthermore, the mutual speaker information (depicted in Stage 3 in Figure 1) among the speaker embeddings are analysed and presented.
3. Contrary to the previous claims [16], we show that acoustic embeddings can be re-synthesised to intelligible audio recordings irrespective of certain types of convolution or feed-forward layers in network architectures of the models.

2. Flexible Gradient Reversal Speaker Anonymization (FleGReSA)

The proposed framework with ASR model training (Stage 1), and evaluation phases (Stage 2 & 3) are shown in Figure 1.

10.21437/Interspeech.2023-1902

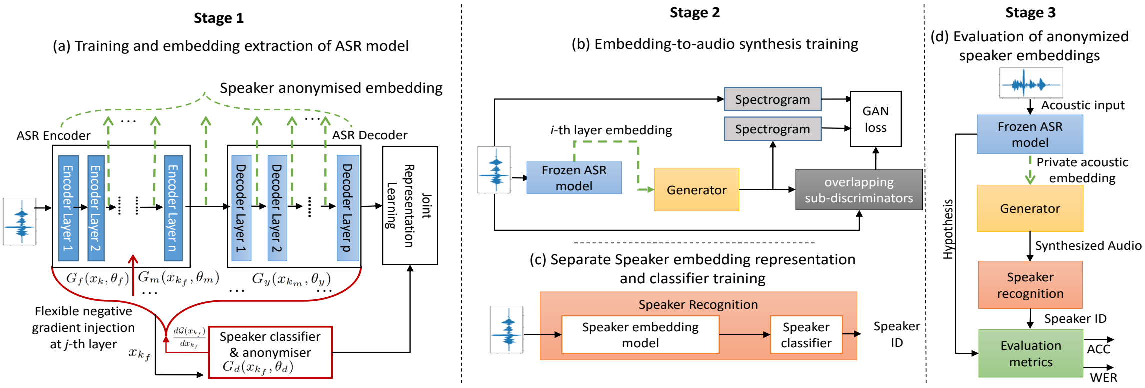


Figure 1: *Our framework proposed for speaker anonymization and evaluation with acoustic embeddings.*

2.1. Stage 1 - Training models and Extracting Embeddings

2.1.1. Training ASR Models

The ASR model training is shown in Figure 1a. We used the conformer model [17] as the baseline ASR model consisting of conformer blocks. A conformer block consists of layer normalisation, Feed-forward, Multi-Headed Self-Attention and Convolution modules [17]. An x-vector [18] speaker classification model is used with the ASR model for speaker anonymization through speaker adversarial training as described below.

2.1.2. Speaker Adversarial Training (SAT)

The SAT aims to learn speaker invariant representations at different layers, and removes speaker specific information from acoustic embeddings [9, 12]. We add gradient reversal layer at different hierarchies of the encoder, with relevant gradient scaling, and make the number of speaker invariant layers flexible. The gradient reversal is a 'pseudo function' [9] $\mathcal{G}(\cdot)$, which defines (a) forward and (b) backward pass with input x_{k_f} by

$$\text{(a)} \quad \mathcal{G}(x_{k_f}) = x_{k_f} \quad \text{and} \quad \text{(b)} \quad \frac{d\mathcal{G}(x_{k_f})}{dx_{k_f}} = -\alpha \cdot \mathbf{I} \quad (1)$$

where x_{k_f} is the output of the i^{th} layer where the gradient reversal $\mathcal{G}(\cdot)$ is applied, α is the gradient scaling factor, and $\mathbf{I}$ is the identity matrix. In the forward-pass (a), it follows the identity transformation, and in the backward-pass (b), it is multiplied by $-\alpha$. When gradient reversal is added at the i^{th} encoder block, the ASR model is split in: (1) feature extractor $x_{k_f} = G_f(x_k, \theta_f)$ which comprises the 1^{st} to the i^{th} ASR encoder block; (2) speaker invariant encoder $x_{k_m} = G_m(x_{k_f}, \theta_m)$ defined by the remaining layers in the ASR encoder; and (3) ASR decoder $G_y(x_{k_m}, \theta_y)$. The k^{th} input sample to the ASR model is x_k, and θ_f, θ_m and θ_y denote the parameters in the feature extractor, speaker invariant encoder and decoder, respectively.

The discriminative speaker classifier $G_d(x_{k_f}, \theta_d)$, which is used to enforce invariant representations, takes input $\mathcal{G}(x_{k_f})$ where θ_d denotes its parameters. The ASR model loss L_y and the speaker classifier model loss L_d are defined by [19–21].

$$L_y\left(\theta_f, \theta_m, \theta_y\right) = L_y(G_y(G_m(G_f(x_k, \theta_f), \theta_m), \theta_y), y_k) \quad (2)$$

$$L_d\left(\theta_f, \theta_d\right) = L_d\left(G_d\left(G_f\left(x_k, \theta_f\right), \theta_d\right), s_k\right) \quad (3)$$

where y_k and s_k are the transcription label and speaker label for the k^{th} sample, respectively. Hence, the final loss is

$$L(\theta_f, \theta_m, \theta_y, \theta_d) = \frac{1}{K}\sum_{k=1}^{K} L_y^k(\theta_f, \theta_m, \theta_y) + \frac{1}{K}\sum_{k=1}^{K} \lambda \cdot L_d^k(\theta_f, \theta_d) \quad (4)$$

where the total number of samples is K and λ is speaker loss regularizer. The gradient of the loss with respect to the input can be written by (dropping arguments of the losses for clarity)

$$\frac{\partial L}{\partial x_k} = \frac{\partial x_{k_f}}{\partial x_k} \cdot \frac{\partial L}{\partial x_{k_f}} = \frac{\partial x_{k_f}}{\partial x_k} \cdot \left(\frac{\partial L_y}{\partial x_{k_f}} + \frac{d\mathcal{G}(x_{k_f})}{dx_{k_f}} \cdot \frac{\partial L_d}{\partial \mathcal{G}(x_{k_f})}\right)$$
$$= \frac{\partial x_{k_f}}{\partial x_k} \cdot \left(\frac{\partial L_y}{\partial x_{k_f}} - \boldsymbol{\alpha}\frac{\boldsymbol{\partial L_d}}{\boldsymbol{\partial x_{k_f}}}\right). \quad (5)$$

where the term in bold is the gradient injected for speaker adversarial training. The speaker classifier used in the speaker adversarial training is based on x-vector [22] model. Unlike the previous works [10, 11], the speaker adversarial classifier is not a pre-trained model and it is trained jointly with the ASR model (Stage 1). After training, the speaker adversarial classifier is removed from the ASR model where the layers are trained to have speaker invariant acoustic representations, and only θ_f, θ_m, and θ_y are used for decoding.

2.2. Stage 2 - Training Embedding-to-audio Synthesis and Speaker Recognition Models

2.2.1. Neural Embedding to Speech Synthesis

Contrary to the previous methods [7, 8], where a voice conversion approach is used to convert the audio to a different speakers voice, we directly anonymize the acoustic embeddings from the ASR model. The speaker privacy in the anonymized acoustic embedding is evaluated using speaker classifiers [10].

The *discriminative* speaker classifiers may be sensitive to small changes (e.g., perturbation difference) in embedding spaces among different ASR models [23, 24]. Moreover, the same utterances may have different embeddings obtained from different ASR models. Therefore, comparing embeddings provided by different ASR models to achieve speaker privacy is not practical. As a result, an extra stage is added to be able to listen to the audio synthesized from acoustic embeddings.

We propose a method to employ acoustic embeddings for audio synthesis and evaluate the anonymization of the generated acoustic embeddings (Stage 2 in Figure 1b). The embedding-audio synthesis model is based on HiFi GAN, and a mixture of multi-period and multi-scale sub-discriminators [25, 26]. During inference, it takes embeddings from different layers and produces high resolution audio synthesis. If x_{k_i} is the acoustic embedding obtained at the ith layer for waveform input x_k, then the synthesised output of the generator $\hat{x}_{k_i} = G_{syn}(x_{k_i})$ has the same dimension as x_k. According to [26], the training loss for the embedding-audio synthesis training is the summation of generator loss L_{mel} and sub-discriminator loss L_{FM} given by

$$L_{mel} = \frac{1}{K}\sum_{k=1}^{K} ||\phi(x_k) - \phi(\hat{x}_{k_i})||_1, \tag{6}$$

$$L_{FM}(D) = \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{T}\frac{1}{N_i}||D^i(x_k) - D^i(\hat{x}_{k_i})||_1 \tag{7}$$

where ϕ is the function used for calculating spectrogram, T is the total number of layers in a discriminator and N_i is the feature dimension of the ith layer output denoted by D^i. Mel-spectrogram loss (L_{mel}) and feature matching loss (L_{FM}) calculate ℓ_1 distances between spectrograms and those between discriminator outputs during training. The discriminator in the synthesis module comprises Q sub-discriminators $\{D_q\}_{q=1}^{Q}$ which are used in the final losses:

$$L_{G_{syn}} = \sum_{q=1}^{Q}\left[\frac{1}{K}\sum_{k=1}^{K}\left[(D_q(x_k)-1)^2 + (D_q(\hat{x}_{k_i}))^2\right] + \lambda_{FM}L_{FM}(D_q)\right] + \lambda_{mel}L_{mel}, \tag{8}$$

$$L_D = \sum_{q=1}^{Q}\frac{1}{K}\sum_{k=1}^{K}\left[(D_q(\hat{x}_{k_i})-1)^2\right] \tag{9}$$

where λ_{mel} and λ_{FM} are loss scaling parameters.

2.2.2. *Training Speaker Embedding and Identification Models*

An x-vector model [22] pre-trained on Voxceleb1 [27] and Voxceleb2 [28] is fine-tuned on LibriSpeech data for learning speaker representations. This model shown in Figure 1c is only used to evaluate the synthesized acoustic embeddings for speaker identification performance.

2.3. Stage 3 - Speaker Anonymization Evaluation

Speaker anonymization is evaluated on the waveforms synthesized from the acoustic embeddings using the generator described in Section 2.2.1 as depicted in Figure 1d. The acoustic embeddings are obtained from different layers of the ASR model and the waverforms are evaluated with the fine-tuned x-vector model (Section 2.2.2).

3. Experimental Setup

3.1. Experimental Setup

Data: The publicly available LibriSpeech [29] corpus has been used for ASR model training (in Figure 1a), embedding extraction (in Figure 1a) and embedding-audio synthesis (in Figure 1b). The train-clean-100 (100 hours) split has been used for training. The *dev-clean*, *test-clean*, and *test-other* splits have been used for validation and testing. Additionally, we have combined train-clean-100 and train-clean-360 into train-clean-460 (460 hours clean speech). This combined set has been used for the training of embedding-to-audio synthesis.

For the speaker adversarial ASR training, the labels for the ASR and speaker classifier models are necessary. The speaker classifier model requires same speakers for training and evaluation. Therefore, before the training, some utterances have been randomly selected and separated from training data for each speaker to create (*test-adv*). The speaker classifier shown in Figure 1c is fine-tuned with dev-clean, 70% of the speaker, leaving 30% for evaluation (*dev-clean-te*).

Train Data	**GRL**	α/λ	**test-adv WER (%)**	**dev-clean WER (%)**	**test-clean WER (%)**	**test-other WER (%)**
train clean 100	**-**	**-**	**-**	**6.14**	**6.18**	**16.28**
	CE3	0.01/0.5	4.55	5.38	5.58	16.27
		0.1/0.3	3.23	5.48	5.72	15.94
		1.5/0.3	3.48	6.19	6.77	18.09
		1.0/0.1	3.51	6.06	6.47	18.27
		0.5/0.5	3.08	5.76	6.23	17.63
		1.0/0.05	3.85	5.68	5.92	17.57
	CE5	0.5/0.5	2.81	6.27	6.74	18.31
		0.1/0.3	3.73	5.34	5.75	15.96
		0.01/0.3	3.69	**5.26**	**5.47**	16.04
		0.5/0.05	2.78	5.54	5.80	16.85
		1.0/0.1	3.36	5.89	6.27	17.36
	CE7	1.0/0.05	4.29	5.78	6.11	17.55
		0.05/0.3	3.69	5.41	5.64	**15.92**
	CE10	1.0/0.3	3.59	5.95	6.31	17.46
		0.1/0.3	4.02	5.47	5.88	17.09
		0.5/0.5	4.12	6.12	6.71	17.93
	CD4	0.5/0.3	3.27	5.60	5.94	17.30
		0.5/0.5	4.21	5.63	5.87	17.32
		1.0/0.1	4.12	5.42	6.10	17.02

Table 1: *An analysis of the ASR performance (WER) applying gradient reversal at different layers of the ASR model with different α and λ, where* GRL *denotes the gradient reversal layer.*

Setup: We performed the experiments in three stages. In the first stage (Figure 1a), an ASR model is trained with speaker adversarial loss. In the second stage (Figure 1b), acoustic embeddings are extracted from different layers, and then the embedding-audio GAN model is trained to reconstruct the original audio. The hyperparameters for the GAN training are similar to [26] *V1*. The synthesis models are trained with the clean 460 hours of LibriSpeech data. In the third stage (Figure 1c), the embedding-audio GAN generator is used to synthesize audio from acoustic embeddings to evaluate the speaker anonymity compared to the original audio utterances and baseline. The second and third stages are evaluation stages. The experiments were implemented using Speechbrain [21].

Baseline: A conformer [17] model with 12 encoder and 4 decoder blocks has been used as the ASR baseline model. The model has $13.3M$ trainable parameters and it is decoded with a language model *shallow fusion* [21], beam size 1. The baseline model is used both for training the ASR model and extracting embeddings for audio synthesis. The baseline model embeddings are compared with the FleGReSA embeddings for evaluating their anonymity compared to the original audio samples.

3.2. Evaluation

The ASR model is evaluated using Word Error Rate (WER), and speaker classifier is evaluated using the unweighted accuracy (WA) metric. ASR performance is evaluated with models where gradient reversal layers are applied at their different layers with different scaling α and λ values. The goal is to analyse the impact of gradient reversal, and stabilise ASR training with scaling weights in different layers when gradient reversal is applied. The ASR decoding setup is same as the baseline. The speaker anonymization of the acoustic embeddings obtained from different layers of the ASR model is evaluated using the speaker identification accuracy based on x-vector as mentioned in Section 2.2.2.

4. Results & Discussion

The ASR performance of the speaker adversarial ASR is shown in Table 1 where: CE denotes *conformer encoder*; CD denotes *conformer decoder*; the number following CE or CD is the embedding layer number; α and λ are scaling factors used in Eq.

(1) and (4). Instead of applying the adversarial layer only at the end of the encoder [10], we propose flexible speaker adversarial at various hierarchies of the encoder/decoder model and found ASR performance improvements. The *test-adv* WER shows the ASR performance on utterances which have common speakers with the training data but not common utterances. The other test sets are standard *dev-clean*, *test-clean* and *test-other*. The overall results given in Table 1 show that the ASR performance obtained from speaker adversarial training improves across the test scenarios compared to the baseline (first row). We observe that adding GRL in the lower layers does not decrease the ASR performance. The weight of the gradient reversal layer is crucial in the initial convergence and overall performance of the ASR model 1. The results show that that high values for α and λ prevent the ASR model from converging. Furthermore, the weight of the gradient reversal layer α is also dependant on the layer of the ASR model where the gradient reversal layer is injected, as the linguistic and speaker information are highly entangled at the initial layers of the encoder of the ASR model [13, 14]. The α and λ weights need to be smaller to make the ASR model stable as lower negative speaker gradients distort the sequential linguistic entanglement in the acoustic embeddings, and it loses the linguistic boundary information. As a result, the ASR model mostly predicts blanks and misaligned word sequences.

Next, we analyze how the layers can become speaker invariant after the intersection of gradient reversal layers. In Table 2, the higher the speaker accuracy, the less anonymous the speaker representations are. The results show with adversarial training, the ASR model embeddings are more speaker redundant. The *adv_CE3D* model shows when the gradient reversal is at layer 3 (CE3) and the embedding is extracted from layer 5 (CE5), the acoustic embeddings are more anonymous compared to the acoustic embeddings extracted from layer 3. This suggests that we can control the trade-off between embedding speaker quality and downstream task performance by flexible adversarial training. Thereby, we achieve speaker anonymity in acoustic embeddings without expensive efforts like voice morphing or conversion [3, 6, 7]. Next, we compare the audio waveform reconstructed using the *baseline* model to the original audio waveform. We observe that plenty of speaker information remains in the acoustic embeddings at the convolution and fully-connected layers obtained from the *baseline*.

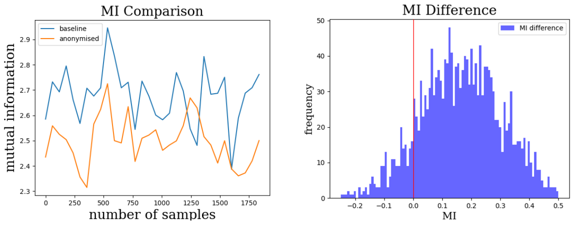
The speaker anonymization of the embeddings is further assessed computing the mutual information (MI) of random variables of the embeddings. For this purpose, we compute the MI using embeddings $\hat{x}^b_{k_i}$ obtained at the i^{th} layer of the baseline model and the embeddings $\hat{x}^a_{k_i}$ obtained at the i^{th} layer of the anonymized model. The MI is computed between the original waveform x_k and the synthesized audio $\hat{x}_{k_i}$ using

$$\mathcal{I}(x_k, \hat{x}_{k_i}) = \sum_{x_k, \hat{x}_{k_i}} p(x_k, \hat{x}_{k_i}) \log \frac{p(x_k, \hat{x}_{k_i})}{p(x_k)p(\hat{x}_{k_i})}. \tag{10}$$

The frequency of the MI difference $\mathcal{I}(x_k, \hat{x}^b_{k_i}) - \mathcal{I}(x_k, \hat{x}^a_{k_i})$ is plotted as a histogram to analyze the information loss among samples in Figure 2. In Figure 2a, the blue line denotes the speaker MIs computed with $\mathcal{I}(x_k, \hat{x}^b_{k_i})$ for dev-clean where $\hat{x}^b_{k_i}$ is generated with the baseline synthesised model (i.e. Baseline in Table 2). The orange line in Figure 2a denotes the speaker MIs calculated as $\mathcal{I}(x_k, \hat{x}^a_{k_i})$ where $\hat{x}^a_{k_i}$ is generated with the anonymized model (i.e. adv_CE3D_v1 in Table 2). The difference between these two curves is displayed in Figure 2b as a histogram. These results show evidence of the speaker information reduction using the anonymized model where a substantial proportion of the utterances is reduced after speaker anonymization. These results corroborate the findings observed in Table 2.

Train Data	Model	GRL	dev-clean WER(%)	α/λ	AE	dev-clean-te SPK-ACC
train clean 100	Original audio	-	-	-	-	96.9
	Baseline	-	6.14	-	CE3	86.6
					CE5	22.1
					CE7	6.4
	adv_CE1	CE1	5.58	0.5/0.05	CE3	73.9
					CE4	41.7
					CE5	42.9
	adv_CE3D_v1	CE3	5.76	0.5/0.5	CE3	**64.6**
					CE4	**33.8**
					CE5	**14.8**
	adv_CE3D_v2	CE3	6.06	1.0 / 0.1	CE3	71.1
					CE5	18.0
	adv_CE5D	CE5	5.54	0.5/0.05	CE5	18.2
					CE6	0.8
					CE7	**0.4**
	adv_CE10D	CE10	5.48	0.5/ 0.3	CE3	68.8
					CE5	70.2

Table 2: *Speaker accuracy on the re-synthesised waveforms from acoustic embeddings at different layers, where* AE *denotes the acoustic embedding extraction point and* SPK-Acc *denotes the unweighted speaker accuracy (%).*



(a) (b)

Figure 2: *(a) Comparison of MI computed using speaker embeddings obtained from baseline and FleGReSA. (b) Frequency of difference of the MI.*

5. Conclusion

In this paper, a flexible gradient reversal speaker anonymization (FleGReSA) and evaluation framework is presented. One of the main benefits of the proposed framework is performing anonymization as an integral part of the ASR model. Once we train the ASR model with the domain adversarial speaker classifier, the latter is discarded. The ASR model is solely employed to provide the anonymous acoustic embeddings. We showed that the training is flexible depending upon the acoustic embedding extraction layer and desired downstream task. The results show that the ASR model is stable and performs better with the adversarial training, while providing significant speaker anonymization on the acoustic embeddings. Experimental results obtained using the LibriSpeech indicate that in the best case the proposed approach achieves a remarkable reduction in speaker recognition accuracy by an absolute 22%. Furthermore, the best ASR performance among the models improves the relative WER of the ASR model by 14%. Furthermore, we have presented an embedding to audio high-quality waveform synthesis model not only comparing speaker information but subjectively listening to the synthesized audio of layer-wise embeddings.

6. References

[1] O. Gupta and R. Raskar, “Distributed learning of deep neural network over multiple agents,” *Journal of Network and Computer Applications*, vol. 116, pp. 1–8, 2018.

[2] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, “Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release,” in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[3] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X. Li, “Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity,” *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, 2018.

[4] 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 *2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2020, pp. 2802–2806.

[5] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, “Enhancing speech privacy with slicing,” in *Interspeech*, 09 2022, pp. 5025–5029.

[6] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, “Differentially private speaker anonymization,” *ArXiv*, vol. abs/2202.11823, 2022.

[7] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Tilli, and N. T. Vu, “Speaker anonymization with phonetic intermediate representations,” *arXiv preprint arXiv:2207.04834*, 2022.

[8] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, “Anonymizing speech with generative adversarial networks to preserve speaker privacy,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 912–919.

[9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[10] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, “Privacy-preserving adversarial representation learning in asr: Reality or illusion?” *ArXiv*, vol. abs/1911.04913, 2019.

[11] M. Jaiswal and E. M. Provost, “Privacy enhanced multimodal neural representations for emotion recognition,” in *AAAI Conference on Artificial Intelligence*, 2019.

[12] W. Zhou, H. Wu, J. Xu, M. Zeineldeen, C. Lüscher, R. Schlüter, and H. Ney, “Enhancing and adversarial: Improve asr with speaker labels,” *arXiv preprint arXiv:2211.06369*, 2022.

[13] Y. K. Singla, J. Shah, C. Chen, and R. R. Shah, “What do audio transformers hear? probing their representations for language delivery & structure,” in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2022, pp. 910–925.

[14] A. Ollerenshaw, M. A. Jalal, and T. Hain, “Insights on Neural Representations for End-to-End Speech Recognition,” in *Interspeech 2021*. ISCA-International Speech Communication Association, 2021, pp. 4079–4083.

[15] A. Ollerenshaw, M. A. Jalal, and T. Hain, “Probing statistical representations for end-to-end asr,” 2022. [Online]. Available: https://arxiv.org/abs/2211.01993

[16] L. Comanducci, P. Bestagini, M. Tagliasacchi, A. Sarti, and S. Tubaro, “Reconstructing speech from cnn embeddings,” *IEEE Signal Processing Letters*, vol. 28, pp. 952–956, 2021.

[17] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” *ArXiv*, vol. abs/2005.08100, 2020.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[19] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” ser. ICML ’06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[20] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, “State-of-the-art speech recognition with sequence-to-sequence models,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, p. 4774–4778.

[21] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018, pp. 5329–5333.

[23] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *stat*, vol. 1050, p. 20, 2015.

[25] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[26] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033.

[27] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” *arXiv preprint arXiv:1706.08612*, 2017.

[28] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.