Towards Privacy-Preserving Speech Representation for Client-Side Data Sharing

Minh Tran, Mohammad Soleymani
Department of Computer Science, University of Southern California, USA
{minhntra, msoleyma}@usc.edu

Abstract
Privacy and security are major concerns when sharing and collecting speech data for cloud services such as automatic speech recognition (ASR) and speech emotion recognition (SER). Existing solutions for client-side privacy mainly focus on voice conversion or voice modification to convert a raw utterance into another one with similar content but different, or no, identity-related information. However, an alternative approach to share speech data under the form of privacy-preserving representations has been largely under-explored. To fill this gap, we propose a speech anonymization framework that provides formal privacy guarantees via noise perturbation to a selected subset of the high-utility representations extracted using a pre-trained speech encoder. The subset is chosen with a Transformer-based privacy-risk saliency estimator. We validate our framework on four tasks, namely, Automatic Speaker Verification (ASV), ASR, SER and Intent Classification (IC) for privacy and utility assessment. Experimental results show that our approach is able to achieve a competitive, or even better, utility compared to the baselines that use voice conversion and voice modification, providing the same level of privacy. Moreover, the easily-controlled amount of perturbation allows our framework to have a flexible range of privacy-utility trade-offs without re-training any components.

Index Terms: privacy, speech representations, speech data release

1. Introduction
Collecting and sharing speech data from local devices is essential for cloud-based machine learning models. From the service providers’ perspective, gaining access to more data enables them to train more robust models and provide better services. From the users’ perspective, sharing personal data allows experiencing more personalized services such as healthcare [1] or automatic speech recognition (ASR) [2]. However, privacy and security are major concerns when sharing or collecting personal speech data, as raw utterances often contain sensitive identifiable information such as voiceprint, racial background, gender, or even health condition. One promising line of solution focuses on client-side privacy [3], providing users full control to protect their privacy by sanitizing raw utterances locally before uploading the data for cloud-based processing without the need for a trusted server.

Several methods attempting to anonymize speakers while preserving the contents have been proposed, including voice modification [4,5], voice conversion [6], and adversarial learning [7]. The majority of these methods rely on sharing data in the form of audio, which is not required for machine learning. On the other hand, pre-trained self-supervised speech representation models such as wav2vec2 [8] and audio HuBERT [9] have become increasingly popular due to their robustness in a wide range of downstream tasks. In fact, many state-of-the-art systems [9,10] often start from the extracted features of pre-trained models. However, because pre-trained speech models are generally trained on natural voices [11], sharing speech data in the form of synthesized or modified audio could potentially limit the capability of the pre-trained models for downstream tasks.

These facts naturally lead to two questions: Is it beneficial to locally process speech data before sharing in the form of privacy-preserving speech representations to leverage the robustness of pre-trained models? Users often have different levels of privacy versus utility expectations. However, pre-training speech representation models with different privacy settings are too costly. Is it possible to sanitize the extracted representations from pre-trained speech models directly, and avoid re-training any components for different levels of privacy? To address these questions, we propose a speech representations anonymization framework. At the core of our approach is a Privacy-risk Saliency Map Estimator that learns to predict the importance of individual representation positions for a speaker identification system. Based on the estimations, the top $k\%$ positions with the highest estimated privacy-risk would be perturbed by adding easily-controlled Laplacian noise, which provides formal privacy guarantees. Figure 1 shows an overview of our proposed framework. We validate our method on four tasks, namely, Automatic Speaker Verification (ASV) for privacy evaluation, Automatic Speaker Verification (ASR), Emotion Recognition (ER), and Intent Classification (IC) for utility performance. Our work provides preliminary evidence that using speech data in the form of extracted representations results in systems achieving competitive performance compared to the commonly-used voice conversion/modification approaches. Moreover, our approach enables a flexible privacy-utility trade-offs.
2. Method

2.1. Privacy-risk Saliency Map Dataset

The idea of saliency maps was originally proposed in computer vision to identify the most salient regions in images for visual concept detection and image understanding. Simonyan et al. [12] introduces Vanilla Gradient, which constructs the saliency map for a given image as the absolute value of the gradient of the input pixels with respect to the loss for the true class of the image. This can be computed via a backpropagation on the class score $S$, output by an image classification model. Smilkov et al. [13] proposed SmoothGrad, which improves upon Vanilla Gradient, by first generating multiple noisy versions of the images by adding randomly sampled noises. SmoothGrad then computes the saliency map for each version of the image using Vanilla Gradient, and output the average of the computed saliency maps. SmoothGrad is able to generate sharper and more accurate saliency maps than Vanilla Gradient.

For a given 2D representation $x \in \mathbb{R}^{n \times d}$ of an utterance, we hypothesize that some positions in $x$ contain more sensitive information than the others (i.e., an adversary can probe more identity-related information from these positions than the others). Hence, we adapt SmoothGrad to find these positions. Specifically, we first train a Speaker Identification (SID) model $M$ on a dataset $D = \{(x_i, y_i)\}$, with $x_i$ being the extracted features from a pre-trained model (e.g., wav2vec2) and $y_i$ being the identity of $x_i$. For each sample $(x_i, y_i)$ in $D$, we compute the privacy-risk saliency map $s_i$ with SmoothGrad, which is the average of $\|M(x_i + z) - M(x_i)\|_2$, where $L_{y_i}$ is the loss of $M$ on the true class $y_i$ of $x_i$ with different versions of noise $z$. As a result, we construct a privacy-risk saliency map dataset $D' = \{(x_i, s_i)\}$ containing the same number of samples as $D$ with $x_i$ and $s_i$ of the same sizes. This dataset is then used to train our Privacy-risk Saliency Estimator.

2.2. Privacy-risk Saliency Estimator

The purpose of the Privacy-risk Saliency Estimator (PSE) is to take in the extracted feature representations $x_i$ for an utterance $u$ and output an estimate for the privacy-risk saliency map $s'_i$ such that $s'_i \approx s_i$. We use the Transformer encoder [14] as our architecture for the PSE. As a high-level overview, the transformer encoder is a stack of Transformer encoder layers, each of which consists of a self-attention module followed by a feed-forward neural network. For a given $x_i \in \mathbb{R}^{T \times d}$, we expect the self-attention modules to capture information along the temporal dimension while the feed-forward neural networks capture important signals from the feature dimension, which enables the architecture to make precise position-level predictions. The model’s parameters are optimized with the $L1$ Loss between the outputs of PSE and the ground-truth saliency maps extracted with a trained Speaker Identification model, explained in detail in Section 2.3.

2.3. Extracted Representations Sanitization

2.3.1. Differential Privacy and the Laplace mechanism

Dwork et al. [15] introduce the concept of Differential Privacy, which is often regarded as the golden standard for privacy analysis. Formally, let $A : X \rightarrow O$ be a randomized algorithm mapping a dataset $X$ to an output $O$. The algorithm $A$ is said to be $\epsilon$-differentially private if for all adjacent inputs $x_1, x_2 \in X$ differing by at most one record, and every output $o \in O$,

$$Pr[A(x_1) = o] \leq e^\epsilon Pr[A(x_2) = o]$$

(1)

The parameter $\epsilon$ quantitatively captures the privacy budget of the algorithm $A$. Specifically, $\epsilon = 0$ implies that $A$ provides perfect privacy while $\epsilon \rightarrow \infty$ implies that $A$ has no privacy guarantees. Any post-processing applied to the output of an $\epsilon$-DP algorithm remains $\epsilon$-DP, which guarantees the robustness of the privatized outputs of $A$ to adversarial attacks.

The Laplace mechanism is a commonly used DP algorithm for local data sanitization [15]. Formally, let $f(x) = x_s \in \mathbb{R}^n$, and $A$ be an algorithm that injects Laplacian noise to $x_s$ such that

$$A(x_s) = x_s + z$$

(2)

where $z = \{z_1, z_2, \ldots, z_n\}$ with $z_i \sim \text{Lap}(\frac{\Delta}{\epsilon})$. $\Delta$ is often called the global sensitivity of $f$, defined as

$$\Delta = \max_{x_1, x_2} ||f(x_1) - f(x_2)||_1$$

(3)

for any adjacent inputs $x_1$ and $x_2$ differing by at most one record. $A$ has been shown to be $\epsilon$-DP.

2.3.2. Sanitization for extracted speech representations

We assume the extracted representations from wav2vec2 can be separated into either sensitive or insensitive positions. In this study, we define the group of sensitive positions as positions with the top $k\%$ highest privacy-risk values estimated by PSE. To improve the utility of the output representations, we only apply the Laplace mechanism to the sensitive positions, i.e., the vector $x_s$ in Equation [2]. We later investigate the effect of $k$ on the privacy and utility of the sanitized outputs. It is important to note that the sensitivity $\Delta$ needs to be determined for the application of the Laplace mechanism. However, it is difficult to find the true global sensitivity of $x_s$ given that little is known about the boundaries of the extracted representations from the speech encoder, i.e., wav2vec2 in this work. To address this issue, we follow Shokri et al. [13] to set a fixed input-independent bounds of $[-1, 1]$ to the sensitive positions, resulting in $\Delta = 2$.

3. Experimental Setups

In this study, we focus on sanitizing the extracted representations from wav2vec2 [8]. However, the proposed framework should be applicable to other pre-trained speech encoders. Our implementations and training procedures follow the SUPERB benchmark [17] to enhance reproducibility. Unlike prior studies that focuses on ASR for utility evaluation, we intentionally pick three downstream tasks with emphasis on different aspects of speech, namely, ASR for content, ER for paralinguistics, and IC for semantics, for a more comprehensive evaluation. Our code is publicly available[4].

3.1. Tasks, Datasets & Models

VoxCeleb1 (SID & ASV). We use the VoxCeleb1 dataset [13] for training the SID and ASV models. It contains approximately 150,000 utterances from 1,211 speakers collected from around 21,819 videos on YouTube. For a fair privacy evaluation, we first split the dataset into two partitions. The first subset contains approximately 50,000 utterances from 400 speakers, and is used to train the SID model and construct the Privacy-risk Saliency Map dataset. The second subset contains around

[4] https://github.com/mtran14/dp_w2v2.git
100,000 utterances from the remaining 811 speakers, which is used to train the ASV model for privacy evaluation. The output representations from wav2vec2 is fed to the SID model consisting of a mean-pooling layer followed by a fully-connected layer to make the speaker identity predictions. The SID model is trained using a cross-entropy loss. For ASV, we train the x-vector model [19] with an AMSoftmax loss [20]. Following prior work [21, 8, 4], we report the equal error rate (EER) as our privacy metric (high EER implies good privacy-preserving representations).

IEMOCAP (ER). We use the IEMOCAP dataset [22] for training the ER model. The original dataset contains approximately 12 hours of data from 10 actors performing improvised or scripted scenarios that are designed to evoke emotions. Following the SUPERB benchmark’s setting, we only keep four emotion classes (neutral, happy, sad, angry) from the IEMOCAP dataset to get an approximately balanced dataset. The ER model adds a mean-pooling layer followed by a fully-connected layer to the encoder for emotion recognition. We report the classification accuracy as the utility metric.

Fluent Speech Commands (IC). We use the Fluent Speech Commands dataset [23] to train our IC model. The dataset contains around 30,000 utterances from 97 speakers with 31 unique intents. Similar to ER, the IC model contains a mean-pooling layer followed by a fully-connected layer. We report the classification accuracy as an utility metric.

LibriSpeech (ASR). We use the commonly used LibriSpeech audiobook dataset [11] to train our ASR model. Specifically, we use LibriSpeech’s train-clean-100/dev-clean/test-clean subsets to train, validate and test our model. The training set contains more than 100 hours of transcribed speech from 251 speakers while both the validation and test set contains more than 5 hours of transcribed speech from 40 speakers. Our ASR model is a 2-layer 1024-unit Bidirectional LSTM, trained with the CTC loss [24] on characters. We do not use any language model to improve the model’s performance during inference. The Word Error Rate (WER) is used as the utility metric.

3.2. Privacy-risk Saliency Estimator implementations

Our PSE is a standard Transformer encoder, which consists of 6 layers with a dropout rate of 0.1. Each transformer encoder layer contains a self-attention module with 12 heads. The sizes of the feed-forward layers in each transformer encoder block are 3072. The model is optimized with the L1 loss using the Adam optimizer with a learning rate of $1e^{-4}$ and a batch size of 32 for 60 epochs. Before training, we split the Privacy-risk Saliency Map dataset into a training and validation sets of approximately 45,000 and 5,000 samples, respectively. We perform model selection based on the lowest loss on the validation set to avoid over-fitting.

3.3. Baseline models

Existing literature on local speech anonymization generally follows two lines: voice conversion to change the speaker embeddings while retaining the content (i.e., transforming the source voice to a target voice or target speaker) [25, 3, 26, 27] or voice modification (i.e., apply different signal processing techniques directly to the source voice to get a perturbed output voice) [4, 28, 5]. In this study, we select two voice conversion methods, namely the Cycle-GAN [29] and the Assem-VC [30], as the baselines for the first line of anonymization approach. For the second line, we use the signal processing-based methods from Kai et al. [4] as our baseline.

Cycle-GAN. The method learns voice conversion via adversarial learning through a two-stage process. In the first stage, an autoencoder is trained to separate speaker-independent features and speaker embeddings with the help of a speaker classifier. In the second stage, the decoder takes as inputs the extracted speaker-independent latent representations from the encoder along with a target speaker embedding to generate a voice of the target speaker without changing the content of the utterance. The model is trained using the CSTR VCTK Corpus [31] with 20 target speakers.

Assem-VC. The model consists of three main components: a linguistic encoder, an intonation encoder, and a decoder. It takes as inputs the raw utterances along with the corresponding transcripts (extracted with an ASR system). The Cotatron linguistic encoder [32] learns to estimate the alignments between the Mel Spectrograms and the transcripts. The linguistic features are then concatenated with the normalized fundamental frequency (F0), followed by a non-causal decoder and HiFi-GAN Vocoder [33] to generate the output voices. The model is trained using the CSTR VCTK [31] and LibriTTS Corpus [34] with 231 target speakers.

R+MS. This approach is a combination of Resampling and Modulation Spectrum Smoothing [15]. The input raw utterance is first stretched and resampled to a different sampling frequency from the original speech without changing the duration. The utterance is then transformed with the modulation spectrum smoothing method that removes the temporal fluctuation of the complex spectrogram obtained from STFT. R+MS is the authors’ recommended method from their voice modification module with the best privacy-utility trade-offs.

R+MS+M+CH+CL. This approach is an extension of R+MS by applying three other voice modification methods to the modified utterance, namely the McAdams Transformation [50], Clipping, and Chorus. The McAdams Transformation method aims at changing the resonant frequencies of the utterance. The Clipping method limits the extremum (maximum and minimum) values in a speech waveform while Chorus slightly modifies the pitch of the utterance. R+MS+M+CH+CL is the most complex method proposed by Kai et al. [4], which gives the best level of privacy protection for the modified utterances.

3.4. Privacy-Utility evaluation setups

We report the Equal Error Rate (EER) from ASV as our privacy metric. As our utility metrics, we report the Word Error Rate from ASR, emotion classification accuracy from ER, and intent classification accuracy from IC. For each task, we train a single model on the extracted features of wav2vec2 on the original utterances from the corresponding dataset, and use the trained model for assessing the privacy-utility performances of both our framework and the baselines. We report the performance of these models in the original row in Table 1.

For the Voice Conversion baselines, we use the trained voice conversion models provided by the official implementations. We use two methods to choose the target speaker for a given utterance. Following Justin et al. [25], our first method converts all speakers’ voices to that of a specific speaker. For the second method, we convert each voice to a random target speaker from the pool of speakers in the trained voice conversion models. This target-speaker selection method has been theoretically shown to provide the highest level of privacy protection [25]. For the Voice modification baseline models, we use the pre-optimized hyper-parameters on the LibriSpeech dataset from the official implementation.
Table 1: Baseline models performance

| Model                  | EER ↑ | IC Acc (%) ↑ | ER Acc (%) ↑ | ASR WER ↓ |
|------------------------|-------|--------------|--------------|-----------|
| Original               | 0.063 | 92.67        | 64.81        | 6.54      |
| Cycle-GAN (one)        | 0.339 | 48.91        | 34.12        | 59.66     |
| Cycle-GAN (random)     | 0.475 | 42.92        | 32.36        | 65.67     |
| Assem-VC (one)         | 0.459 | 50.51        | 38.26        | 44.86     |
| Assem-VC (random)      | 0.487 | 46.04        | 32.94        | 57.09     |
| R-MS                   | 0.216 | 82.78        | 48.49        | 8.52      |
| R-MS-M-CH-C            | 0.454 | 20.01        | 31.18        | 77.12     |

Figure 2: Proposed framework utility performance. Dashed lines represent the “random perturbed positions” baselines.

Figure 3: Proposed framework privacy performance. Dashed lines represent the “random perturbed positions” baselines.

4. Results

Table 1 shows the privacy-utility evaluation for the baseline models. Figure 2 and 3 show the evaluation results for the utility and privacy metrics, respectively. To highlight our PSE’s ability to correctly identify positions with sensitive information, we add another baseline that randomly chooses k% of the positions in the extracted wav2vec2 representations for perturbation with similar settings to our approach.

We can observe that the proposed approach enables a very flexible range of privacy-utility trade-offs, with EER ranging from 0.13 to 0.50 while the utility metrics ranging from random guessing performance to approximately the original performance. On the higher EER range, the approach shows competitive, if not better, performance compared to the baselines. For example, compared to Assem-VC (random) with 0.487 EER, perturbing 20% of the extracted representations with $\epsilon = 0.5$ yields a similar EER of 0.49 yet better utility metrics with an ER Accuracy of 43.3%, IC Accuracy of 45.3%, and ASR WER of 49.9. The proposed framework also consistently outperforms the baseline models on the Emotion Recognition benchmark. In comparison with the best performing baseline on ER (R-MS) with 0.216 EER and 48.5% accuracy, our approach can achieve an accuracy of 52.2% with 0.287 EER when 20% of the extracted representation is perturbed with $\epsilon = 4.0$. On the higher EER range, Assem-VC(one) achieves ER Accuracy of 38.3% with EER of 0.459 while our method is able to get ER Accuracy of 44.2% with EER of 0.478 (20% perturbed with $\epsilon = 1.0$).

From Figure 3 we can see that our method achieves superior (higher) EER curves compared to the random perturbation baselines in all $\epsilon$ and perturbation ratio variations. It is also important to note the gap between our method and the random perturbation baselines gets smaller as the proportion of perturbed positions $k$ gets larger, with the most significant gap observed at $k = 20\%$. These observations suggest that the PSE is able to select high privacy-risk positions in the wav2vec2 extracted representations for perturbations. However, from Figure 2 we can see that the dashed curves generally have better utility performances compared to the solid curves, excluding Speech Recognition with a reversed effect. This implies that the selected positions from PSE might potentially be high-utility positions for downstream tasks.

5. Conclusion

In this study, we propose a framework to perturb the extracted representations from pre-trained speech models, with the goal of generating privacy-preserving representation for data sharing. The framework consists of three stages: generating a privacy-risk saliency map dataset for representations extracted from a selected pre-trained model, training a Privacy-risk Saliency Estimator (PSE), and perturbing the extracted representations based on the selected positions from PSE. We validate our framework on a wide range of downstream tasks and find that sharing speech data in the form of extracted representations is competitive to the existing approaches that share speech via voice conversion or voice modification. We hope the results motivate further exploration in the direction of privacy-preserving pre-trained representation learning.
6. References

[1] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, “Fedhealth: A federated transfer learning framework for wearable healthcare,” IEEE Intelligent Systems, vol. 35, no. 4, pp. 83–93, 2020.

[2] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Abharrd, H. Suk, A. Graemstein, P. Beaufays et al., “Personalized speech recognition on mobile devices,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5955–5959.

[3] P. Wu, P. P. Liang, R. Salakhutdinov, and L.-P. Morency, “Understanding the tradeoffs in client-side privacy for speech recognition,” arXiv preprint arXiv:2101.08919, 2021.

[4] H. Kai, S. Takamichi, S. Shiota, and H. Kiya, “Lightweight voice anonymization based on data-driven optimization of cascaded voice modification modules,” in 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2021, pp. 560–566.

[5] K. Hashimoto, J. Yamagishi, and I. Echizen, “Privacy-preserving sound to degrade automatic speaker verification performance,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5300–5304.

[6] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, “Speaker de-identification via voice transformation,” in 2009 IEEE Workshop on Automatic Speech Recognition & Understanding. IEEE, 2009, pp. 529–533.

[7] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, “Privacy-preserving adversarial representation learning in asr: Reality or illusion?” in INTERSPEECH, 2019.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.

[10] L.-W. Chen and A. Rudnicky, “Exploring wav2vec 2.0 fine-tuning for improved speech emotion recognition,” arXiv preprint arXiv:2110.06309, 2021.

[11] 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). IEEE, 2015, pp. 5206–5210.

[12] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” arXiv preprint arXiv:1312.6034, 2013.

[13] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “Smoothgrad: removing noise by adding noise,” arXiv preprint arXiv:1706.03825, 2017.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[15] C. Dwork, A. Roth et al., “The algorithmic foundations of differential privacy,” Found. Trends Theor. Comput. Sci., vol. 9, no. 3–4, pp. 211–407, 2014.

[16] R. Shokri and V. Shmatikov, “Privacy-preserving deep learning,” in Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 2015, pp. 1310–1321.

[17] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-T. J. Lai, K. Lakhota, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin et al., “Superb: Speech processing universal performance benchmark,” arXiv preprint arXiv:2105.01051, 2021.

[18] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” Proc. Interspeech 2017, pp. 2516–2620, 2017.

[19] 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). IEEE, 2018, pp. 5329–5333.

[20] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin softmax for face verification,” IEEE Signal Processing Letters, vol. 25, no. 7, pp. 926–930, 2018.

[21] N. Tomashenko, B. 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,” 2020.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” Language resources and evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[23] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, “Speech model pre-training for end-to-end spoken language understanding,” Proc. Interspeech 2019, pp. 814–818, 2019.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connexionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.

[25] 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). IEEE, 2020, pp. 1–6.

[26] T. Justin, V. Širuč, S. Dobrišek, B. Vesnivec, I. Ipišič, and F. Mišelšek, “Speaker de-identification using diphone recognition and speech synthesis,” in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 4. IEEE, 2015, pp. 1–7.

[27] 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,” in INTERSPEECH 2020, 2020.

[28] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, “Speaker anonymisation using the mcdams coefficient,” in INTERSPEECH 2021. ISCA, 2021, pp. 1099–1103.

[29] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-s. Lee, “Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations,” Proc. Interspeech 2018, pp. 501–505, 2018.

[30] K.-w. Kim, S.-w. Park, J. Lee, and M.-c. Joe, “Assem-vc: Realistic voice conversion by assembling modern speech synthesis techniques,” arXiv preprint arXiv:2104.00931, 2021.

[31] J. Yamagishi, C. Veaux, K. MacDonald et al., “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92),” 2019.

[32] S.-w. Park, D.-y. Kim, and M.-c. Joe, “Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data,” arXiv preprint arXiv:2005.03295, 2020.

[33] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” Advances in Neural Information Processing Systems, vol. 33, pp. 17022–17033, 2020.

[34] H. Zou, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen, “Libritts: A corpus derived from librispeech for text-to-speech,” in INTERSPEECH, 2019. [Online]. Available: https://arxiv.org/abs/1904.02882

[35] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, “The naist text-to-speech system for the blizzard challenge 2015,” in Proc. Blizzard Challenge workshop, vol. 2. Berlin, Germany, 2015.

[36] S. E. McAdams, Spectral fusion, spectral parsing and the formation of auditory images. Stanford university, 1984.