Interspeech 2024
1-5 September 2024, Kos, Greece

Prosody-Driven Privacy-Preserving Dementia Detection

Dominika Woszczyk[1], *Ranya Aloufi*[1,2], *Soteris Demetriou*[1]

[1]Imperial College London, UK
[2]Taibah University, Saudi Arabia

d.woszczyk19@imperial.ac.uk, r.aloufi18@imperial.ac.uk, s.demetriou@imperial.ac.uk

Abstract

Speaker embeddings extracted from voice recordings have been proven valuable for dementia detection. However, by their nature, these embeddings contain identifiable information which raises privacy concerns. In this work, we aim to anonymize embeddings while preserving the diagnostic utility for dementia detection. Previous studies rely on adversarial learning and models trained on the target attribute and struggle in limited-resource settings. We propose a novel approach that leverages domain knowledge to disentangle prosody features relevant to dementia from speaker embeddings without relying on a dementia classifier. Our experiments show the effectiveness of our approach in preserving speaker privacy (speaker recognition F1-score .01%) while maintaining high dementia detection score F1-score of 74% on the ADReSS dataset. Our results are also on par with a more constrained classifier-dependent system on ADReSSo (.01% and .66%), and have no impact on synthesized speech naturalness.[1][2]

Index Terms: privacy, speech verification, dementia

1. Introduction

Problem Statement. Advances in deep learning, combined with non-invasive biomarkers like speech, offer a promising opportunity for large-scale disease diagnosis. Researchers have investigated the use of speech signals for detecting different medical conditions, such as neurodegenerative diseases (e.g., Parkinson's and Alzheimer's) [1] and respiratory ailments (e.g., COVID-19) [2].

Speaker embeddings (e.g., i-vector, x-vector, and ECAPA-TDNN) are a type of feature set utilized in the detection of early signs of diseases like dementia [3]. However, speaker embeddings often contain more information than required for their intended tasks, which poses potential privacy concerns. For example, they contain speaker-specific information, which makes them effective in automatic speaker verification (ASV) and zero-shot text-to-speech (TTS) among others. This unintentional information leakage might violate GDPR policies [4] and its data minimization principle [5] (i.e., "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed"), leaving individuals vulnerable to discrimination, extortion, and targeted advertisements by third–parties. Moreover, since speaker embeddings can be used to predict both dementia and verify a speaker, they can be potentially classified as "individually identifiable health information". Under the HIPAA Privacy Rule [6] such information is defined as "protected health information" (PHI) and must be properly de-identified.

Prior Works. Anonymization or de-identification refers to the task of concealing the speaker's identity while retaining the linguistic content, thereby making the data unlinkable [7]. According to the ISO/IEC International Standard 24745 on biometric information protection [8], biometric references must be irreversible and unlinkable for full privacy protection. Most of the proposed works focus on protecting speaker identity, using voice conversion (VC) mechanisms [9, 10]. Beyond speaker identity, various works propose to protect speakers' attributes and paralinguistic information such as emotion [11, 12], gender [13], age [12] or nationality [14]. Attribute obfuscation allows a speaker to conceal specific personal aspects in their voice representation while still maintaining overall performance [15, 13]. Adversarial training disentangles dimensions in latent spaces for speaker verification while minimizing detection of specific attributes [11]. The need of an external attribute classifier, especially for low-resourced attributes like dementia with no or limited training data, is a significant constraint. An alternative approach is to work at the feature level rather than the utterance level [13, 16, 17]. By extracting and sanitizing feature representations from speech, we can share privacy-aware features instead of complete utterances. Noé et. al., in [13], for example, proposed a Normalizing Flow-based architecture to disentangle sex information in x-vectors.

Our Approach. In this work, we leverage prosody disentanglement as a method for speaker anonymization in dementia detection. Specifically, our goal is to preserve the dementia attribute in speaker embeddings while reducing speaker-related information (identity). Often, adversarial training or multi-task learning is utilized to confine information within bottlenecks, separating out dimensions such as content, pitch, rhythm, and timbre [18, 10]. However, we take a different approach, focusing on prosodic features that are known to be prominent in dementia speech, such as articulation rate, pauses, and disfluencies. Our hypothesis is that by disentangling these features from speaker representations, we can effectively obscure the speaker's identity while minimizing any impact on dementia-related information. Our system achieves this without relying on dedicated classifiers, using domain knowledge and adversarial learning techniques.

Below we summarize the **main contributions** of this work:

1. **Novel Approach**: We propose a novel method for preserving privacy in speaker embeddings in low-resource settings through domain knowledge and prosody disentanglement.
2. **New Application Domain** We shed light on a sensitive attribute, dementia, that has not been extensively investigated in previous attribute obfuscation and anonymization works.

[1]The code is available at https://github.com/domiwk/privacy-preserving-ad-detection

[2]Samples are available at https://shorturl.at/cNS39

10.21437/Interspeech.2024-2137

2. Privacy-Preserving Dementia Detection

2.1. Threat Model

Our threat model focuses on medical speech-processing systems developed for detecting dementia and their handling. These systems use audio data to generate embeddings that can help identify signs of dementia. This raises significant privacy concerns as speaker embeddings can potentially reveal individuals' identities. We consider an adversary that has access to the anonymized embeddings and aims to re-identify the user by using the embeddings not for their primary purpose (dementia classification) but for speaker recognition. This information could be exploited for discriminatory purposes or targeted advertising. Effective anonymization techniques should prevent such linkage attacks while preserving speech naturalness, intelligibility, and performance in dementia detection.

We aim to safeguard the privacy of user identity in different scenarios that involve voice embeddings or medical speech processing. This involves obfuscating any identifying information that a user may not want to share, without compromising the functionality of the system. We also emphasize the need to offer various privacy settings to balance the trade-off between privacy and utility, and to encourage transparent privacy management practices.

2.2. Proposed Approach

We devise a prosody-based privacy-preserving extraction for speaker representations, trained on a larger auxiliary dataset that does not need to have the target attribute label. We leverage domain knowledge for the task of dementia detection and propose a method that performs disentanglement that focuses on prosody features relevant to dementia. Indeed, previous studies have shown that features such as speech rate, mean energy levels, number of pauses and lengths [19, 20] are informative for dementia classification. We explore two approaches: adversarial learning and mutual information-guided shuffling.

2.3. Adversarial Learning for Speaker-Prosody Disentanglement

The proposed adversarial model is based on domain adversarial training [21] which aims to create domain-invariant latent spaces by maximising the domain discriminator's confusion while minimizing the task-specific loss. In our case, we train our model to extract dementia-relevant prosody features while maximising the loss of speaker-relevant features. To train this model, we can take advantage of a larger dataset and extract prosodic features. At inference time, we run the model on any dataset we wish to anonymize. The model consists of several components: the feature extractor that extracts speaker representations, and prosody regressors for each prosodic feature. The loss of our model is defined as:

$$\mathcal{L} = \sum_{i=0}^{\mathrm{n}} \mathcal{L}_{\mathrm{ADpros}_i} - \sum_{j=0}^{\mathrm{n}} \lambda_j \mathcal{L}_{\mathrm{SPKpros}_j} \tag{1}$$

where $\mathcal{L}_{\mathrm{ADpros}_i}$ denotes the i^{th} loss associated with the i^{th} dementia-relevant prosody feature extraction, and $\mathcal{L}_{\mathrm{SPKpros}_j}$ the j^{th} loss associated with the j^{th} speaker-relevant prosody feature regressor. λ_j is a hyperparameter that controls the balance between the different objectives and n is the total number of prosody regressors.

2.4. Mutual Information-Guided Shuffling

We propose a feature selection approach based on mutual information to identify important features relevant to dementia while perturbing less relevant features to lower speaker recognition accuracy. Our method is similar to [16] where they extract Shapley values from a classifier to select key dimensions, however, unlike theirs, our approach is *classifier-free*. The intuition behind this approach is that critical features associated with dementia are likely to have higher mutual information with the target variable (dementia), while shuffling the remainder to preserve dementia-related features but decrease speaker recognition accuracy. We also explore using prosody features instead of the dementia label as the target variable.

Mutual information is a measure of mutual dependence between two random variables. In our context, we compute the mutual information between the distribution of dimension of the speaker embeddings and the distribution of the dementia label across the dataset. Formally, the mutual information $I(X;Y)$ between random variables X and Y is defined as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{2}$$

where $p(x,y)$ is the joint probability mass function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability mass functions of X and Y, respectively. To estimate the mutual information, we use the k-Nearest-Neighbour-based MI estimator [22] which can be applied to both continuous and discrete variables. We design the feature selection strategy as follows: 1) Compute the mutual information between each embedding dimension and the dementia/prosody variable across the whole corpus. 2) Select top n dimensions as important dementia-related features. 3) Shuffle the remaining features. When combining several features, compute the top n dimensions for each and compute the union as the set of important features.

3. Experimental Setup

3.1. Datasets

We train our disentanglement model on the LibriSpeech dataset [23], a corpus of read English speech. We selected the `train-clean-100` subset which consists of 100 hours of speech from 251 speakers. We split the data into 25685, and 2850 samples for train, and dev sets respectively. We extract prosody features for each sample.

For testing dementia detection, we use two publicly available datasets widely studied in dementia classification: ADReSS [24] and ADReSSo [24]. The ADReSS (ADR) dataset is a subset of the DementiaBank dataset, a collection of recordings from control (CD) and dementia (AD) patients describing the Cookie Theft Picture [25]. Manual transcripts are provided and we split each sample into segments using the sentence-level timestamps from the transcripts. We group the original test and train sets and split the data to contain one sample per speaker in the test set and keep the remaining samples as training data. We end up with 1723 samples (868 CC | 855 AD) in the training set and 156 (78 CC | 78 AD) in the test and validation sets. The ADReSSo (ADRo) dataset is another subset of DementiaBank designed for detecting dementia from spontaneous speech only, without access to manual transcriptions. The original set consists of 151 train samples (87 CC | 74 AD) and 71 test samples (35 CC | 35 AD). We use the provided segmentation timestamps

to isolate segments spoken by the patients and get 2705 samples in the train set (920 CC | 1026 AD) and 231 samples (74 CC | 87 AD) in the test and validation sets, one per speaker.

3.1.1. Data Processing

For the dementia datasets, we split samples into segments as described in Section 3.1. For both LibriSpeech and Dementia datasets, we extract a series of articulation features, number and length of pauses, f0 and mean energy. We compute the features with Parselmouth and Praat scripts made available by Feinberg [26] and normalise them to be within [0,1]. When extracting embeddings, we trim the segments to 30s.

3.2. Implementation Details

We select a pre-trained ECAPA-TDNN [27] embedding extractor, regarded as the state-of-the-art (SOTA) for speaker embeddings. We use the SpeechBrain [28] implementation and pretrained model which was trained on the VoxCeleb dataset. We augment the ECAPA-TDNN embedding extractor with classifiers for each prosodic feature, each consisting of two layers with 126 hidden dimensions. We use the Mean-Square-Error Loss for the prosody regressors and investigate λ values in $\{1,5,10,30\}$ and set λ to 1.0. We train the model with a batch size of 8 with cyclical learning rate policy (CLR) and set the base and maximum learning rates to 1e-7 and 1e-5 respectively. We finetune the models for 15 epochs with early stopping. We compare our model to several models: a model trained on a dementia dataset to classify dementia while fooling the speaker classifier (ADV SPK$_{AD}$ + AD), a model trained on the LibriSpeech dataset to fool the speaker classifier (ADV SPK$_{LS}$), a random shuffling method (Shuffle$_{Random}$) and a shuffling from [16] where we train an XGBoost model [29] on dementia and select top n Shapley values (Shuffle$_{Shap}$). For the mutual information shuffling method (Shuffle$_{MI}$), we compute the mutual information using the sklearn library [30]. The dementia classifier is a two-layer model with a hidden space size of 8 and a ReLu activation between layers, trained with Binary Cross-Entropy. The speaker classifier is a two-layer model with a hidden space size of 96 trained with Cross-Entropy Loss. We optimise regressors with bayesian search.

3.3. Evaluation Metrics

Privacy. We measure the privacy gain through the drop of an adversary's (speaker classifier from Section 3.2) F1-score and speaker verification Equal Error Rate (EER) score (using cosine distance). Due to the size of our dataset, we focus on a black-box setting and do not adapt our adversary. We evaluate the anonymized embeddings against an SVM with a radial basis function kernel, as it performed the best on both datasets.

Utility. To evaluate the ability to detect dementia, we train a neural dementia classifier on top of the embeddings and report the F1-score. We implement a feed-forward network with two layers, a hidden space size of 8 and a ReLu activation between layers. Additionally, we perform zero-shot speech generation with the new embeddings and measure MOSNet [41], SI-SDR [42], and STOI [43], which are used as objective metrics to evaluate the quality of speech signals, and WER (Word Error Rate) with Whisper [31]. These metrics are used to assess the quality of the processed signal compared to the original signal. We pick YourTTS [32] and SpeechT5 [33], two open-source zero-shot TTS systems that have shown SOTA performance.

4. Results

4.1. Prosodic Features Selection

We choose prosody features based on mutual information (MI) with dementia labels in the ADReSS dataset, selecting the top 35 quantiles. Then, we assess their importance in speaker recognition by computing MI with speaker labels. Figure 1 shows MI scores for features w.r.t. class and speaker identity. Notably, mean f0 and energy are significant for both dementia and speakers. However, mean energy is strongly linked with speaker identity, while f0 is a known speaker characteristic. To remove speaker information, we decide to separate features into prosodic and disfluency features to disentangle speakers and dementia. We select speech rate (spr), number of pauses (pnum), length of pauses (plength) and number of syllables as dementia features and mean f0 (f0) and energy (nrg) as speaker-relevant for the ADV PROS and Shuffle$_{MI_pros}$ systems.

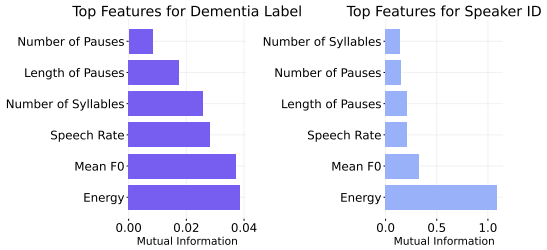

Figure 1: *Mutual information scores of key prosodic features extracted from audio segments for dementia label (left) and speaker identity (ID) (right) on the ADReSS dataset.*

4.2. Overall Results

Table 1: *Comparison of anonymization systems' performance. We report the dementia detection F1-score (AD), speaker recognition F1-score (SPK), their 95% confidence intervals and Equal Error Rate (EER).*

System	ADReSS			ADReSSo		
	AD ↑	SPK↓	EER (%)↑	AD ↑	SPK↓	EER (%) ↑
Original	.81 ± .06	.33 ± .06	35	.73 ± .06	.36 ± .04	35
ADV SPK$_{AD}$ + AD	.64 ± .06	.00 ± .02	47	.63 ± .06	.00 ± .02	45
ADV SPK$_{LS}$	.80 ± .13	.69 ± .01	15	.75 ± .1	.75 ± .01	16
ADV$_{nrg}$ PROS	.73 ± .13	.01 ± .01	43	.66 ± .10	.01 ± .01	44
Shuffle$_{Random}$	.63 ± .08	.01 ± .00	41	.61 ± .06	.02 ± .04	41
Shuffle$_{Shap}$	.62 ± .07	.02 ± .02	41	.64 ± .07	.04 ± .02	40
Shuffle$_{MI_AD}$	.63 ± .07	.02 ± .02	41	.62 ± .06	.02 ± .01	42
Shuffle$_{MI_pros}$	.62 ± .08	.01 ± .02	44	.68 ± .06	.02 ± .01	43

Table 1 shows the performance of the anonymization systems. We see that ADV SPK$_{AD}$ + AD drops the speaker F1-score to 0% for both ADR and ADRo and increases the EER by 12%, 10% respectively, giving the best anonymization performance. We note that the EER on the original samples is quite high (35%), which we attribute to the noisy nature of the dataset. However, in this work, we are interested in the relative increase. We see the system reduces the AD detection F1-score to 65%, on par with the shuffling approaches. Through experimentations, we observe that the energy (nrg) gave us stronger privacy guarantees across subsets, as shown in Table 3. We report ADV$_{nrg}$ PROS as our best system. ADV$_{nrg}$ PROS, drops the accuracy to 0.01% and adds 8% to the EER while preserving 73% F1-score on ADR, a drop of 7% from the original performance, making it the best-performing approach. On ADRo it

Table 2: *Zero-Shot TTS Quality Metrics and WER on the ADReSS dataset.*

TTS	System	MOSNet↑	SI-SDR↑	STOI↑	Avg. Rank↓	WER (%)↓
SpeechT5	Original	2.84	-69.01	.26	3.33	23
	ADV SPK_{AD} + AD	2.84	-73.41	.23	1.83	30
	ADV_{nrg} PROS	2.84	-72.23	.21	2.83	35
	$Shuffle_{MI\_pros}$	2.84	-70.73	.26	2.50	21
	$Shuffle_{Shap}$	2.84	-72.16	.23	2.83	34
YourTTS	Original	2.77	-68.57	.24	2.33	27
	ADV SPK_{AD} + AD	2.82	-68.60	.23	1.33	30
	ADV_{nrg} PROS	2.82	-67.90	.23	2.83	29
	$Shuffle_{MI\_pros}$	2.82	-68.43	.23	2.00	27
	$Shuffle_{Shap}$	2.82	-68.23	.23	2.66	28

achieves comparable performance to ADV SPK_{AD} + AD and shuffling approaches. We compare it to the model trained adversarially against a speaker classifier ADV SPK_{LS}. Surprisingly, the models improved the speaker recognition and dementia classifiers on both datasets. We evaluated it in combination with other features, and a similar pattern was seen, hence we report only ADV SPK_{LS}. Finally, we evaluate different shuffling approaches. We report the systems selecting top 50 features as it gave us the best privacy/utility tradeoff for all systems. We note that the selection process with MI based on the label ($Shuffle_{MI\_AD}$) and prosody ($Shuffle_{MI\_pros}$) seems to add little to the accuracy when compared to $Shuffle_{Random}$. Nevertheless, the shuffling approaches achieve similar results to the strong baseline ADV SPK_{AD} + AD, while being simpler.

4.3. Zero-shot TTS-based Evaluation

The Voice Privacy Challenge (VPC) revealed that all methods, including x-vector embeddings and signal processing techniques, could lead to a degradation in speech naturalness and intelligibility [7, 34]. We thus conducted an objective evaluation on the ADReSS dataset using speech synthesis models using two text-to-speech (TTS) systems: SpeechT5 and YourTTS (zero-shot TTS systems) to assess the effectiveness of anonymized embeddings in text-to-speech (TTS) tasks. We utilize sentence-level transcriptions as input data and condition the TTS system on speaker embeddings. For a fair comparison, we regenerate the raw recordings using the raw embedding extracted from the original recordings. We then use the anonymized embeddings of different settings to generate different test sets. Rankings were based on average performance across all metrics, with lower average ranks indicating better overall speech quality. MOSNet represents the Mean Opinion Score predicted by a neural network model for the quality of speech signals. Higher MOSNet, SI-SDR, and STOI scores indicate better quality and speech intelligibility. Table 2 reports the results of our proposed method under different settings and the baseline (synthesized raw recordings) using two TTS systems. We find that the anonymization results in almost the same voice distinctiveness as the data originally had, and, according to speech recognition, produces intelligible speech recordings. ADV SPK_{AD} + AD seems to strike the best balance in maintaining voice distinctiveness while ensuring anonymity, as shown by its lower Avg. Rank, even though all methods introduce some level of degradation in SI-SDR and WER, which is a trade-off for achieving speaker anonymity.

4.4. Ablation study

We investigate the impact of dementia-relevant and speaker-relevant prosodic features and report the results in Table 3. We find that the addition of prosody features improves dementia detection with a negligible impact on privacy metrics. Nearly all features are useful for preserving dementia, while the speaking rate did not seem to affect the detection.

Table 3: *Different prosody sets. We report the dementia detection F1-score (AD), speaker recognition F1-score (SPK), their 95% confidence intervals, and Equal Error Rate (EER).*

System	ADReSS			ADReSSo		
	AD↑	SPK↓	EER (%)↑	AD↑	SPK↓	EER (%)↑
ADV_{f0}	.79 ± .13	.17 ± .02	29	.72 ± .11	.20 ± .01	29
ADV_{f0} PROS	.76 ± .13	.29 ± .02	29	.74 ± .11	.27 ± .02	29
ADV_{nrg_f0}	.73 ± .13	.03 ± .02	32	.72 ± .11	.03 ± .01	33
ADV_{nrg_f0} PROS	.74 ± .13	.02 ± .02	32	.68 ± .11	.03 ± .01	33
ADV_{nrg}	.65 ± .11	.01 ± .01	45	.67 ± .11	.01 ± .01	46
ADV_{nrg} PROS	.73 ± .13	.01 ± .01	43	.66 ± .11	.01 ± .01	44
ADV_{nrg} $PROS_{no\_nsyll}$	.68 ± .13	.01 ± .01	43	.66 ± .11	.01 ± .01	45
ADV_{nrg} $PROS_{no\_plength}$	.70 ± .13	.01 ± .01	44	.64 ± .11	.00 ± .01	47
ADV_{nrg} $PROS_{no\_pmum}$	.69 ± .13	.00 ± .01	44	.61 ± .11	.01 ± .01	46
ADV_{nrg} $PROS_{no\_spr}$	.71 ± .13	.02 ± .02	43	.64 ± .11	.01 ± .01	44

5. Discussion

In this work, we show that we can successfully anonymize speaker embeddings and preserve dementia detection by disentangling prosodic features relevant to dementia. Albeit the original EER is quite high, its increase demonstrates the effectiveness of our approach. Future work will explore a stronger adversary in a white-box setting and more robust to challenging speech. Furthemore, the approach can be extended to other attributes or health conditions. Nevertheless, a limitation of our work is the domain knowledge and feature analysis required to fine-tune the disentanglement. For our use case, the mean energy was an important feature in removing speaker information. However, this is dataset-specific and the generalizability of our system would need to be evaluated on a larger dataset, and explore more features. In terms of feature selection on the embeddings, future work would investigate explainable methods to understand what embeddings encode throughout the extraction process to better extract/obfuscate information. Finally, dementia has an important impact on linguistics and their content which could be further explored to preserve dementia better. However, this also raises other privacy concerns where patient transcripts might be leaked and calls for content disentanglement as well. We leave the exploration of this problem to future work.

6. Conclusion

We present a novel approach for anonymizing speaker embeddings for privacy-preserving dementia detection. Our experiments show the potential of using prosody disentanglement to isolate specific attributes from speaker identity in low-resource settings. By training an embedding extraction model against prosodic extractors on an auxiliary dataset, we effectively minimise speaker identity information while preserving the diagnostic utility for dementia. We show the effectiveness of shuffling selected embedding dimensions for anonymization with an impact similar to classifier-dependent adversarial learning. Notably, our disentanglement approach outperforms the traditional adversarial training, showcasing the importance of the dataset size. Among our selected features, disfluency and articulatory features contribute to preserving dementia with little impact on the speaker's privacy. Our work is a first step towards privacy-preserving speaker embeddings for healthcare applications and we hope to inspire further refinements of these approaches.

7. References

[1] S. de la Fuente Garcia, F. Haider, and S. Luz, "Cross-corpus feature learning between spontaneous monologue and dialogue for automatic classification of alzheimer's dementia speech," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5851–5855.

[2] E. Casanova, A. Candido Jr, R. C. F. Junior, M. Finger, L. R. S. Gris, M. A. Ponti, and D. P. P. Da Silva, "Transfer learning and data augmentation techniques to the covid-19 identification tasks in compare 2021." in *Interspeech*, 2021, pp. 446–450.

[3] C. Botelho, T. Schultz, A. Abad, and I. Trancoso, "Challenges of using longitudinal and cross-domain corpora on studies of pathological speech." 2022.

[4] eur lex.europa.eu, "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)." [Online]. Available: https://tinyurl.com/6tk3j9aw

[5] commission.europa.eu, "What personal data is considered sensitive?" [Online]. Available: https://tinyurl.com/2fmu22j6

[6] hhs.gov, "The hipaa privacy rule," accessed online on March 2024. [Online]. Available: https://tinyurl.com/yhbpsame

[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," in *INTERSPEECH 2020*, 2020.

[8] ISO/IEC, "Iso/iec 24745: Information security, cybersecurity and privacy protection — biometric information protection." May 2021.

[9] B. M. L. 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)*. IEEE, 2020, pp. 2802–2806.

[10] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, and J.-F. Bonastre, "Adversarial disentanglement of speaker representation for attribute-driven privacy preservation," *arXiv preprint arXiv:2012.04454*, 2020.

[11] O. Chouchane, M. Panariello, O. Zari, I. Kerenciler, I. Chihaoui, M. Todisco, and M. Önen, "Differentially private adversarial autoencoder to protect gender in voice biometrics," in *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*, 2023, pp. 127–132.

[12] F. Teixeira, A. Abad, B. Raj, and I. Trancoso, "Privacy-oriented manipulation of speaker representations," *arXiv preprint arXiv:2310.06652*, 2023.

[13] P.-G. Noé, A. Nautsch, D. Matrouf, P.-M. Bousquet, and J.-F. Bonastre, "A bridge between features and evidence for binary attribute-driven perfect privacy," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3094–3098.

[14] C. Luu, S. Renals, and P. Bell, "Investigating the contribution of speaker attributes to speaker separability using disentangled speaker representations," in *Interspeech 2022*. ISCA, 2022, pp. 610–614.

[15] 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.

[16] C. Lavania, S. Das, X. Huang, and K. Han, "Utility-preserving privacy-enabled speech embeddings for emotion detection," 2023.

[17] M. Tran and M. Soleymani, "Privacy-preserving Representation Learning for Speech Understanding," in *Proc. INTERSPEECH 2023*, 2023, pp. 2858–2862.

[18] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[19] V. Vincze, G. Szatlóczki, L. Tóth, G. Gosztolya, M. Pákáski, I. Hoffmann, and J. Kálmán, "Telltale silence: temporal speech parameters discriminate between prodromal dementia and mild alzheimer's disease," *Clinical Linguistics & Phonetics*, vol. 35, no. 8, pp. 727–742, 2021.

[20] P. Pastoriza-Dominguez, I. G. Torre, F. Dieguez-Vide, I. Gómez-Ruiz, S. Geladó, J. Bello-López, A. Ávila-Rivera, J. A. Matias-Guiu, V. Pytel, and A. Hernández-Fernández, "Speech pause distribution as an early marker for alzheimer's disease," *Speech Communication*, vol. 136, pp. 107–117, 2022.

[21] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of machine learning research*, vol. 17, no. 59, pp. 1–35, 2016.

[22] B. C. Ross, "Mutual information between discrete and continuous data sets," *PloS one*, vol. 9, no. 2, p. e87357, 2014.

[23] 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.

[24] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The adress challenge," *arXiv preprint arXiv:2004.06833*, 2020.

[25] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[26] D. R. Feinberg, "Parselmouth praat scripts in python," Jan 2022. [Online]. Available: osf.io/6dwr3

[27] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[28] 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.

[29] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[32] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[33] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li, and F. Wei, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5723–5738.

[34] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Tilli, and N. T. Vu, "Speaker Anonymization with Phonetic Intermediate Representations," in *Proc. Interspeech 2022*, 2022, pp. 4925–4929.