INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Utility-Preserving Privacy-Enabled Speech Embeddings for Emotion Detection

Chandrashekhar Lavania[1], *Sanjiv Das*[2], *Xin Huang*[2], *Kyu J. Han*[1]

[1]AWS AI Labs
[2]Amazon Web Services
{clavania, sanjivda, xinxh, kyujhan}@amazon.com

Abstract

Audio privacy has been undertaken using adversarial task training or adversarial models based on GANs, where the models also suppress scoring of other attributes (e.g., emotion, etc.), but embeddings still retain enough information to bypass speaker privacy. We use methods for feature importance from the explainability literature to modify embeddings from adversarial task training, providing a simple and accurate approach to generating embeddings for preserving speaker privacy while not attenuating utility for related tasks (e.g., emotion recognition). This enables better adherence with privacy regulations around biometrics and voiceprints, while retaining the usefulness of audio representation learning.

Index Terms: adversarial training, speaker identification, emotion scoring, feature shuffling

1. Introduction

Audio (in particular speech) is now an important modality in multimodal machine learning (ML), [1, 2]. Representation learning is a first step, i.e., converting the data of any modality into "embeddings"—fixed size tensors—that are then pipelined into downstream supervised or unsupervised learning models, [3] is an early example.

We introduce a new methodology to generate speech embeddings that are privacy cognizant, i.e., they are optimized for a speech classification task (e.g., emotion scoring, arousal detection, task identification, etc.), while sharply attenuating speaker identification. The approach appears to be simple—fit a seq2seq model [4] using gated recurrence units (GRUs) that emits embeddings using a scheme optimized for emotion scoring and penalized for speaker identification—known as adversarial task training as in [5]. But, we show that this approach does not work as intended, and we propose a method to modify speech embeddings such that they are not usable for speaker identification but remain viable for emotion scoring.

Speech processing presents specific and interesting privacy considerations around biometric data. Biometric information includes retina or iris scans, fingerprints, voiceprints, hand scans, facial geometry, DNA, and other unique biological information. Privacy preserving embeddings produced using the methods in this paper are practical, especially when the intended use case does not call for speaker identification.

We focus on a single speech analysis task, i.e., emotion scoring (analogous to detecting sentiment from text), though the methodology applies to all utility tasks. Speech may be classified into common emotions such as anger, disgust, fear, happiness/joy, sad, and neutral, etc., see [6], which also studies detection of emotion level into low, medium, high, and unspecified. Other models detect valence and arousal [7].

There are various approaches to suppression of speech attributes such as gender, speaker, emotion, intention, etc., [5]. Voice conversion, which entails changing the voice of the speaker to another is widely used [8, 9, 10, 11]. Voice morphing, which alters speech attributes such as pitch and intensity, is also often applied. These methods, while highly effective for speaker masking, also end up masking sensitive emotional states of the speaker, which is not the intention in this paper, which differs from this literature as follows: (i) Instead of masking the emotional state *and* identity of the speakers, we wish to identify the emotion yet suppress speaker identification. Hence, we do not wish to hide all attributes of audio in representations from an adversary as in [12], i.e., we wish to preserve "utility" of the audio dataset (see [13] for an example on tabular data using GANs; [14], [15] for examples on images). (ii) We do not aim to modify the original audio, as in [8], or style transfer applications [11, 9], but specifically the embeddings generated from the audio that may be used for speaker identification, in order to better comply with biometric laws. (iii) Rather than using GANs to generate synthetic audio, as in [16], we directly modify the embeddings by shuffling components using feature importance techniques, an approach that is simple and effective in terms of generating embeddings that sharply attenuate speaker identification while only marginally impacting emotion detection.

The rest of the paper proceeds as follows. In Section 2 we review adversarial speech embeddings. Section 3 discusses the data used in this study. Section 4 presents an analysis and disadvantages of using adversarial embeddings. Section 5 introduces the new embedding shuffling method and its efficacy, and concluding comments are offered in Section 6.

2. Adversarial Audio Embeddings

The adversarial task training approach generates embeddings designed to attenuate speaker identification through an objective function (L) that equals the difference between accuracy on the emotion classification task and on the speaker identification task:

$$L = (w_{emo} \cdot L_{emo}) - (w_{spk} \cdot L_{spk}) \quad (1)$$

Here, L_{emo} (L_{spk}) is the cross-entropy loss on the emotion detection (speaker identification) tasks, respectively, with weights $w_{emo}, w_{spk} \geq 0$, where each weight is $1/(-\log(1/n_j))$, $j = \{emo, spk\}$, and n_j is the number of classes in j.

A sequence-to-sequence model is used to generate the embeddings, which will vary depending on the relative values of w_{emo} and w_{spk}. This model is diagrammed in Figure 1. In related work, [15, 16] build a model based on GANs to suppress specific sensitive attributes in speech data. A similar approach may also be taken to induce fairness in representations, as is done with GANs, usually applying min-max methods as

10.21437/Interspeech.2023-1075

in [12, 13, 17]. Our paper (i) shows that these embeddings do not suppress speaker identification, and (ii) introduces a novel technique to modify these embeddings to suppress speaker identification while supporting other utility tasks. For parsimony, we focus on the emotion detection task, because this is the one that is often suppressed in the literature.

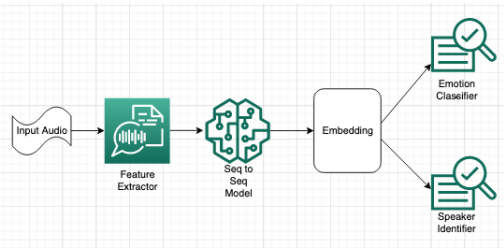

Figure 1: *Emotion classification with discouraged speaker identification. Embeddings from the seq2seq model are passed into task identification layers for both emotion classification and speaker identification and entered into the composite loss function in equation 1.*

3. Data

We use three datasets for assessment of our approach. First, we use the CREMA-D dataset[1] that is labeled for emotion scoring. The data set contains 7,442 speech files from 91 actors (48 male, 43 female, ages 20–74), in both, WAV and MP3 formats. The actors have varied races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Emotions fall into six categories: anger, disgust, fear, happy, neutral and sad. Speech is recorded on 12 sample sentences that were spoken by the actors. Evaluators rate the speech samples and 95% of the files have more than 7 ratings. Since the dataset contains both, emotion and speaker labels, it satisfies the requirement for the specific loss function used in this paper.

Second, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [18]. The dataset of 1440 speech samples is constructed from 24 professional actors (equal female/male), speaking two lexically-matched statements in a neutral North American accent. Labels are: calm, happy, sad, angry, fearful, surprise, and disgust. It has both speech and song, and we only use the speech data, for which there are 22 speakers. The main results are shown on these two widely known datasets.

Third, for robustness we assess the methodology on the audio of quarterly corporate earnings conference calls (ECs). We collected 130 earnings calls audio from six major tech companies: Apple, Amazon, Google, Meta, Microsoft, Netflix, for the period 2019 Q1 through 2022 Q2. We used Amazon Transcribe to diarize the audio and extract the audio for every spoken sentence separately along with the speaker ID, resulting in 46274 sentences across 608 unique earnings call and speaker combinations. We pruned speakers with less than 20 utterances, to make speaker identification easier and make it harder for the new approach to work. Unlike the first two datasets, which are labeled for both emotion and speaker ID, the dataset of ECs does not have emotion labels. However, the purpose of our experiments is to suppress speaker identification, a goal that is not impacted if we have synthetic emotion labels. We create emotion scores using the text transcripts of these earnings calls, using well-established sentiment scoring NLP algorithms. A set of emotion labels is generated using text polarity as a proxy for emotion, scored using an algorithm in Amazon SageMaker [19].[2] The polarity score assigned to each sentence ranges from -1 to $+1$, with several sentence scores at exactly zero, since there may be no positive or negative words in the sentence. Hence, we reduce polarity to three ordinal categories, {-1,0,+1}, using the sign of polarity to establish the label.[3]

[1] https://github.com/CheyneyComputerScience/CREMA-D

4. Embeddings Analysis

4.1. Embeddings Architecture

As shown in Figure 1, (a) the end-to-end framework consists of a feature extractor, a sequence to sequence model, an emotion classifier and a speaker identifier. Inspired by Chung and Zisserman [20], a 6 layer convolutional network is used to extract 512 dimension feature embeddings. The output from the feature extractor is fed into the seq-to-seq model (Figure 1). (b) For CREMA-D, the sequence to sequence model uses a unidirectional GRU [21]. A 2 layer GRU is used with a 512-dimensional hidden state. Furthermore, fixed length snippets are sampled from the entire audio to create a fixed length sequence. A snippet size of 0.5 seconds is used to create a sequence of 3 snippets. A dropout of 0.5 is used during training. The hyperparameters for RAVDESS are the same as those of CREMA-D except the number of layers of the GRU and the snippet size are 3 and 1.2 seconds, respectively. The output of the final unrolling of the seq-to-seq model is used to produce the 512 dimensional embeddings that are fed to the classifier. (c) The emotion classifier is a 4 layer feed forward network with the speaker identifier also being a 4 layer feed forward network.

4.2. Adversarial Training Efficacy

We train seq2seq embeddings in different ways. (i) For emotion detection while suppressing speaker identification, i.e., $w_{emo} > 0, w_{spk} > 0$ (adversarial). (ii) Embeddings from the emotion identification task only (non-adversarial). The last row in Table 1 demonstrates that adversarial training works well as noted in [5]. We drop speaker identification accuracy to close to zero, while maintaining accuracy for emotion scoring. These results are visually corroborated in Figure 2, where emotions cluster cleanly based on the embeddings but speakers do not, for both types of embeddings. The emotion-only embeddings (second row) obviously do better. In fact, when adversarial speaker suppression is implemented (top plots in Figure 2), there is evidence of some speaker identification versus when no speaker suppression is undertaken (lower plots). Therefore, adversarial training does not eliminate speaker identification information from the embeddings, as described in the next subsection.

4.3. Adversarial Training Failure

We verify similar issues (the negative result noted in [5]), i.e., that the embeddings still enable speaker identification when fine-tuned separately from the pipeline in Figure 1. In order to ascertain how well adversarial training eliminates speaker identity from the embeddings, we attempted to fit models for speaker

[2] For documentation on this algorithm, see `https://sagemaker-jumpstart-industry-pack.readthedocs.io/en/latest/smjsindustry.nlp_scorer.html`.

[3] While the first two datasets are in the public domain, the last dataset was collected by downloading and then processed further. This data is in the public domain and is available to anyone for reconstruction.

Table 1: *Accuracy of models on the CREMA-D and RAVDESS datasets. Train/Test splits that have overlapping speakers make it harder to suppress speaker identification*

	CREMA-D Accuracy (%)		RAVDESS Accuracy (%)	
Model	Emo	Spk	Emo	Spk
GRU (emotion only)	60.03	-	72.28	-
GRU (speaker id only)	-	67.97	-	98.60
GRU + Diff in CE Loss	60.60	0.06	69.44	0.00

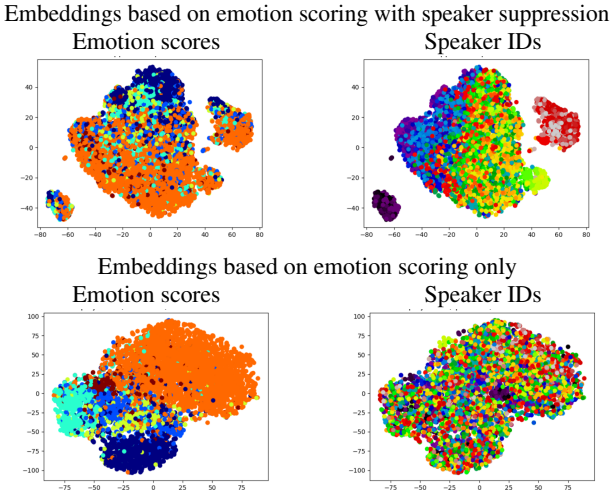

Figure 2: *TSNE projection plots for differentially trained sequence to sequence embeddings (with overlapping speakers in the train and test datasets) on CREMA-D. The left side plots are colored with the emotion labels and the right side plots are colored with speaker IDs. The top pair of images are based on embeddings from the model trained on the loss function in equation 1 where $w_{emo} > 0, w_{spk} > 0$. The bottom pair of images are based on a model where only emotion detection is trained.*

identification using the 512-dimensional embedding feature vector. We used the stack-ensembled model framework in AutoGluon (AG), which has very high accuracy on test datasets (it achieves top 1% leaderboard rankings in Kaggle competitions.[4]) While adversarial training does attenuate speaker identification to some extent—from 68% to 61% on CREMA-D and 98% to 80% on RAVDESS, speaker identification accuracy is still too high to be privacy reassuring. The reason for this is that even though training is adversarial, information from the speaker labels leaks back into the embeddings since they are still being used for adversarial training. Therefore, we propose a simple post-processing step on the embeddings to drastically reduce speaker id accuracy. We denote this as "embedding shuffling" and note that this is very general and may also be applied to embeddings from the emotion scoring task and it does not need embeddings from adversarial training.

[4]See [22], `https://github.com/autogluon/autogluon`

5. Embedding Shuffling

In a feature (or embedding) matrix where we stack 512-dimensional feature (or embedding) vectors into a matrix, we retain the M_E most important feature columns (for emotion detection) and shuffle the remaining columns. If M_E contains any of the top M_S speaker detection features then we shuffle those as well. We vary $M_E = M_S \equiv M = \{75, 50, 25, 15, 10\}$, and as M declines, we expect to see declining accuracy in both emotion detection and speaker identification. If the scheme works well, emotion detection only attenuates slightly whereas speaker identification accuracy drops sharply. Implementation is via the following steps:

1. Fit embeddings (we use dimension 512) to the data using adversarial training (Figure 1).
2. Use these embeddings (a) to fit a *speaker identification* model using any ML model of choice (we used XGBoost); (b) denote as vector f_S the importance-ranked feature dimensions in the 512 embedding vector using any explainability method (we used SHAP [23], the specific version for trees based on [24]).
3. Use the embeddings from (1) to (a) fit an *emotion detection* model using any ML model of choice (we used AG); (b) determine the importance-ranked *feature dimensions* (denoted as vector f_E) of dimension 512 using any explainability method. We used the column permutation method in AG—this method works well when the labels are few (< 10 as is the case with emotion scoring). But with speaker identification, we may have hundreds of speakers, in which case XGBoost+treeSHAP[5] works well as in (2).
4. (a) In the 512 dimensional embedding matrix, keep the top M_E features in the f_E vector fixed and column permute the rest. (b) If there are any features in the top-ranked M_S features in f_S that are in the top M_E features in f_E, permute those as well.
5. Refit both (i) the speaker identification and (ii) emotion detection models with the shuffled embeddings using AG. The results are shown in Figure 3.

5.1. CREMA-D dataset

Based on the procedure above, no embedding shuffling corresponds to 512 fixed features. As shown in Figure 3, first plot, the emotion scoring task has better fit (f1 score) and the speaker suppression task delivers much lower performance, attenuating to less than 10% of the original level, in line with that achieved by [15] for images using GANs, and much lower (i.e., better) than accuracy levels on audio [9] ($\sim 14.2\%$). Hence, training on one task and shuffling using feature importance from both tasks offers a way of suppressing speaker identification without material attenuation of the accuracy of emotion detection. (Similar results obtain when the target metric is accuracy, balanced accuracy, or MCC.) Further, we see that better results are achieved when using the embeddings from emotion scoring only, because speaker id label leakage is circumvented.

5.2. RAVDESS dataset

Figure 3, second plot, shows that feature shuffling attenuates emotion scoring by only 10% but speaker identification f1-score drops by as much as 63%. Both emotion scoring and speaker

[5]`https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Understanding%20Tree%20SHAP%20for%20Simple%20Models.html`

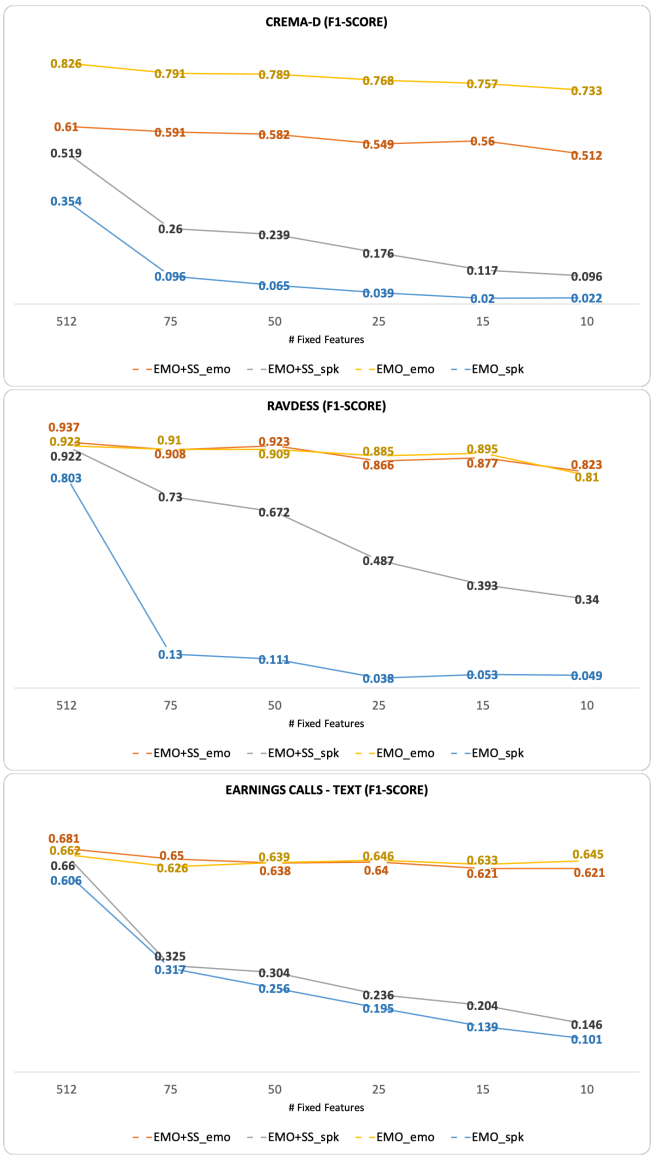

Figure 3: *Suppression of speaker identification after embeddings shuffling. We show the number of fixed (unshuffled embedding dimensions) on the x-axis, ranging from 512 (no shuffling) to 10 (502 features shuffled). (i) $EMO + SS_{emo}$ (dark orange) is F1-score for emotion detection based on adversarial training embeddings; (ii) EMO_{emo} (dark yellow) is for emotion detection w/o adversarial training; (iii) $EMO + SS_{spk}$ (dark grey) is F1-score for speaker identification with adversarial training; (iv) EMO_{spk} (blue) is for speaker identification w/o adversarial training.*

identification metrics are much higher on the RAVDESS dataset than the CREMA-D dataset, despite fewer rows of data. This may be because the quality of the audio is better, and the number of speakers is also far fewer (22 in RAVDESS versus 91 in CREMA-D). The results further confirm that post-processing with emotion task (non-adversarial) embeddings performs better for attenuating speaker identification.

5.3. Earnings Calls dataset

Both the CREMA-D and RAVDESS datasets were specifically prepared for emotion detection using actors. As a robustness exercise, we introduce a third dataset that we constructed from corporate earnings calls by extracting sentences spoken by various speakers on the calls. We label each sentence with the speaker ID and also create synthetic text-based emotion labels, using polarity scoring of the text in each sentence. The number of speakers in this dataset is much larger and none of the speakers is emoting in a directed manner.

The original dataset has 46274 rows and 608 unique earnings call+speaker combinations. Stratified sampling was used to break the dataset into train and test data. Also only speakers who had at least 20 and no more than 80 utterances were retained. After pruning speakers, we get 9955 rows and 216 speakers, i.e., a large multiclass dataset. We undertake a 90:10 train:test split and repeat the analysis in Section 5 using text-based emotion labels. Figure 3, bottom plot, shows that our embeddings shuffling approach barely reduces accuracy on the emotion detection task, but sharply attenuates accuracy on speaker identification. In contrast to the results on the CREMA-D and RAVDESS datasets, where using emotion task only embeddings offers better speaker identity suppression, the EC dataset shows the same sharp speaker suppression for both, emotion task only embeddings and adversarial task training embeddings.

6. Discussion and Conclusion

We show that adversarial task training of embeddings retains sufficient residual information to enable speaker identification, and we find that careful post-processing using embedding shuffling on important features cures this deficiency. We achieve better speaker suppression using shuffling of embeddings trained on the emotion task only, rather than on the embeddings from adversarial task training. (a) We pursued the specific emotion scoring task because speech signals are used in the finance industry to detect positive and negative affect in earnings calls [25]. (b) This new approach is not specific to emotion scoring and generalizes to more than two tasks and to combination with GAN-based methods and other evaluation metrics such as WER. (c) The approach uses fixed length snippets but variable length snippets can be accommodated with the same procedure. (d) The methodology is agnostic to ML training model choice and to explainability technique used for feature importance. These easy to implement privacy approaches will assist in many practical applications.

7. References

[1] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language," *arXiv:2202.03555 [cs]*, Feb. 2022, arXiv: 2202.03555. [Online]. Available: http://arxiv.org/abs/2202.03555

[2] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating Multimodal Information

in Large Pretrained Transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2359–2369. [Online]. Available: https://aclanthology.org/2020.acl-main.214

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[4] A. Shukla, K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Visually Guided Self Supervised Learning of Speech Representations," Feb. 2020, arXiv:2001.04316 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2001.04316

[5] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?" in *Interspeech 2019*, Sep. 2019, pp. 3700–3704, arXiv:1911.04913 [cs]. [Online]. Available: http://arxiv.org/abs/1911.04913

[6] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313618/

[7] A. Baird, E. Parada-Cabaleiro, C. Fraser, S. Hantke, and B. Schuller, "The Perceived Emotion of Isolated Synthetic Audio: The EmoSynth Dataset and Results," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ser. AM'18. New York, NY, USA: Association for Computing Machinery, Sep. 2018, pp. 1–8. [Online]. Available: https://doi.org/10.1145/3243274.3243277

[8] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks," Jun. 2017, arXiv:1704.00849 [cs]. [Online]. Available: http://arxiv.org/abs/1704.00849

[9] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '18. New York, NY, USA: Association for Computing Machinery, Nov. 2018, pp. 82–94. [Online]. Available: https://doi.org/10.1145/3274783.3274855

[10] R. Aloufi, H. Haddadi, and D. Boyle, "Emotionless: Privacy-Preserving Speech Analysis for Voice Assistants," Aug. 2019, arXiv:1908.03632 [cs, eess, stat]. [Online]. Available: http://arxiv.org/abs/1908.03632

[11] M. Pasini, "MelGAN-VC: Voice Conversion and Audio Style Transfer on arbitrarily long samples using Spectrograms," Dec. 2019, arXiv:1910.03713 [cs, eess]. [Online]. Available: http://arxiv.org/abs/1910.03713

[12] H. Edwards and A. J. Storkey, "Censoring Representations with an Adversary," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016. [Online]. Available: http://arxiv.org/abs/1511.05897

[13] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Context-Aware Generative Adversarial Privacy," *Entropy*, vol. 19, no. 12, p. 656, Dec. 2017, number: 12 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1099-4300/19/12/656

[14] Y. Kang and S. Choi, "Learning Features with Structure-Adapting Multi-view Exponential Family Harmoniums," in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2013. [Online]. Available: http://arxiv.org/abs/1301.3539

[15] J. Martinsson, E. L. Zec, D. Gillblad, and O. Mogren, "Adversarial representation learning for synthetic replacement of private attributes," *CoRR*, vol. abs/2006.08039, 2020, arXiv: 2006.08039. [Online]. Available: https://arxiv.org/abs/2006.08039

[16] D. Ericsson, A. Östberg, E. L. Zec, J. Martinsson, and O. Mogren, "Adversarial representation learning for private speech generation," Jun. 2020, arXiv:2006.09114 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2006.09114

[17] C. Huang, P. Kairouz, and L. Sankar, "Generative Adversarial Privacy: A Data-Driven Approach to Information-Theoretic Privacy," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Oct. 2018, pp. 2162–2166, iSSN: 2576-2303.

[18] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, p. e0196391, May 2018, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391

[19] S. R. Das, M. Donini, M. B. Zafar, J. He, and K. Kenthapadi, "FinLex: An effective use of word embeddings for financial lexicon generation," *The Journal of Finance and Data Science*, vol. 8, pp. 1–11, Nov. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405918821000131

[20] J. S. Chung and A. Zisserman, "Signs in time: Encoding human motion as a temporal image," Aug. 2016, arXiv:1608.02059 [cs]. [Online]. Available: http://arxiv.org/abs/1608.02059

[21] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111. [Online]. Available: https://aclanthology.org/W14-4012

[22] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola, "AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data," Mar. 2020, arXiv:2003.06505 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2003.06505

[23] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[24] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s42256-019-0138-9

[25] W. J. Mayew and M. Venkatachalam, "The Power of Voice: Managerial Affective States and Future Firm Performance," *The Journal of Finance*, vol. 67, no. 1, pp. 1–43, 2012, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2011.01705.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2011.01705.x