



An Attention-Based Method for Guiding Attribute-Aligned Speech Representation Learning

Yu-Lin Huang¹, Bo-Hao Su¹, Y.-W. Peter Hong^{1,2}, Chi-Chun Lee^{1,2}

¹Department of Electrical Engineering, National Tsing Hua University, Taiwan

²Institute of Communication Engineering, National Tsing Hua University, Taiwan

huang8592301@gmail.com, borriusu@gapp.nthu.edu.tw, {ywhong, cclee}@ee.nthu.edu.tw

Abstract

The rich personal information contained in speech signal can lead to privacy leakage and unfair prediction for speech based technology. In this work, we propose a feature-scoring variational autoencoder (FS-VAE) to handle these issues by performing attribute alignment for speech representation learning. FS-VAE performs attribute alignment by using attention-based scoring machines guided by two additional penalty terms. After obtaining the attribute-aligned representation, we can then choose and mask the nodes containing specific attribute of interest based on the requirement in the downstream tasks. We evaluate our methods on tasks of PP-SER (identity-free emotion recognition) and PP-SV (emotion-less speaker verification). Our proposed method achieves better utility maintenance and competitive privacy protection compared to the most recent attribute-aligned representation learning method.

Index Terms: speech representation, feature scoring, privacy, fair, attribute alignment

1. Introduction

Speech is the most natural human communication medium that has motivated a thriving effort in the development of speech related technology [1, 2, 3, 4, 5]. The richness of information in human's speech signal [6] while useful but also concerning. As speech signal contains rich personal information [7], e.g., identity, gender, and emotion, users may unexpectedly disclose these sensitive attributes while utilizing speech based services. On the other hand, unfairness may occur as data-driven approaches naturally inherit biases, e.g., gender bias [8] or racial discrimination [9]. As effort being devoted toward achieving trustworthy AI, devising strategy to obtain a speech representation that mitigates biases and privacy concerns directly at the front-end representation is becoming the prevalent approach.

Recently, many works utilize adversarial strategy to eliminate a pre-defined attribute in the speech representation to deal with these issues [10, 11, 12]. While promising, adversarial learning suffers from inflexibility, i.e., one has to retrain an encoder to obtain a specific attribute-removed representation under each setting of an application. Hence, in a recent work, Huang et al. proposed an attribute-aligned speech representation learning method [13], which aligns the task-specific attributes in a particular order for the speech representation, e.g., emotion related (without speaker identity) dimensions are concentrated in the top half of the hidden nodes where speaker identity (without emotion) locate in the bottom half. By aligning attributes along node dimensions, one can flexibly choose and mask the hidden nodes depending on the application scenarios. Specifically, Huang et al. [13] proposed a layered representation variational autoencoder (LR-VAE) to handle a two-attribute scenario (emotion and speaker identity) and demonstrated competitive

privacy-preserving performances with a single encoder (as compared to two for adversarial learning approaches).

However, since LR-VAE depends on manually designed two monotonic dropout functions to align two attributes in the hidden nodes, crafting such functions that extend to three or four or even more attributes can be complicated if not infeasible. In this work, we propose a feature-scoring variational autoencoder (FS-VAE) to achieve attribute alignment. Instead of designing functions to align attributes, we guide the attribute aligned representation learning process by using task-specific attention-based scoring machines along with two additional losses, i.e., attention penalty loss and diversity loss. This approach separates task-specific attributes to distribute distinctively on different latent dimensions. In downstream tasks, one can choose and mask those dimensions needed to maintain a high main-task recognition performance and protect privacy by removing those un-needed dimensions according to the learned attention weights from these scoring machines.

In this work, we present two task definitions: a privacy-preserving speech emotion recognition (PP-SER) that protects identity in SER and a privacy-preserving speaker verification (PP-SV) that protects emotion in SV. We evaluate our method on the MSP-Podcast [14]. Comparing to LR-VAE, the current state-of-the-art attribute-aligned representation learning method, our proposed FS-VAE achieves competitive PP-SER performance (1.86% WFS better, 2.26% EER worse) and improved PP-SV performance (1.08% EER better, 0.30% WFS worse). We also observe that FS-VAE concentrates the task-specific attributes to a fewer number of nodes.

2. Methodology

2.1. Dataset description

We evaluate our method on two tasks, PP-SER and PP-SV. For evaluation, an emotional speech dataset with multiple speakers is required. Hence, we utilize the MSP-Podcast corpus [14], one of the largest emotional corpus with many speakers. In total, the MSP-Podcast contains 33262 speaking turns amounting to 56 hours. In this work, in order to compare fairly to the previous work [13], we perform 5-class emotion classification: neutral, angry, sad, happy and disgust. The distribution of the 5 emotion classes are: neutral: 53.05%, angry: 8.81%, sad: 3.95%, happiness: 27.10%, and disgust: 7.09%. We used the standard splits in Release 1.4 that contains 610 speakers in training set, 30 speakers in development set, and 50 speakers in testing set. Note that the speakers in each set are disjoint.

2.2. Feature extraction

We apply wav2vec2.0 [15], a self-supervised speech representation trained by masking the speech input and solving contrastive task, as input feature. Wav2vec2.0 can be seen as an uni-

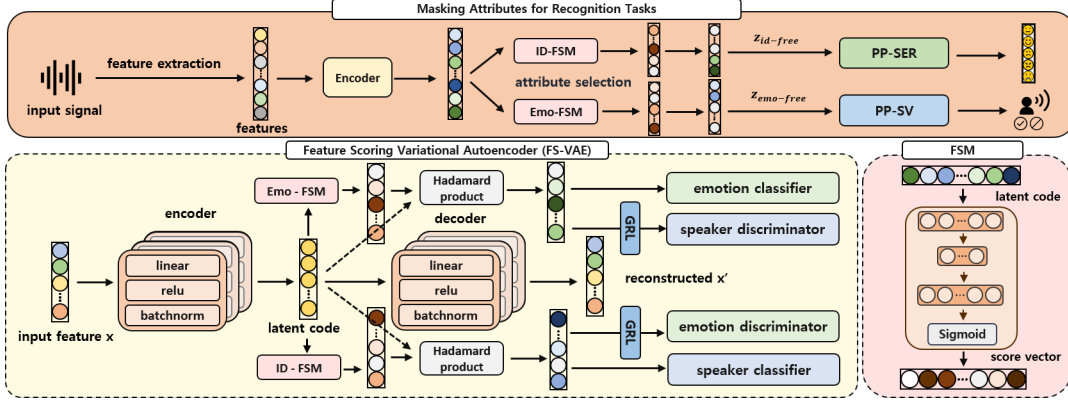


Figure 1: An illustration of our proposed FS-VAE. Note that $Z_{emo-free}$ stands for emotion-free representation, $Z_{id-free}$ stands for identity-free representation, and FSM stands for feature scoring machine.

versal front-end speech embedding and has achieved outstanding results across numerous downstream applications [16, 17]. Specifically, we extract wav2vec2.0 embedding by the released pre-trained model [18], wav2vec2-base, that was trained on the LibriSpeech [19]. Notice that the output of wav2vec2.0 is frame based. We apply average pooling along the time axis to obtain a 768 dimensional feature vector for each utterance.

2.3. Attribute-aligned learning strategy

Attribute-aligned learning strategy aims to learn an encoder which forces task-specific information to condense and distinctively distribute on specific node dimensions. In this work, we propose a feature-scoring variational autoencoder (FS-VAE) to achieve attribute-alignment, including VAE as the representation learning backbone, task-specific feature-attention mechanism, with two designated loss terms, i.e., attention loss and diversity loss.

We apply VAE [20] as backbone for disentangled representation learning, which factorizes the input feature into independent latent dimension. The loss function of VAE is defined as:

$$L_{VAE} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + D_{KL}(q(z|x)||p(z)) \quad (1)$$

Here, $D_{KL}(\cdot)$ stands for the non-negative Kullback-Leibler divergence, which encourages the distribution of the latent dimension to be close to an isotropic Gaussian.

2.3.1. Feature-scoring attention mechanism

To perform attribute-selection in downstream tasks, we require a mechanism that distributes the task-specific information distinctively, i.e., encouraging each node to be responsible for a single attribute. We employ an attention-based mechanism, i.e., feature-scoring machines (FSM) [21, 22], on the latent vector to capture the attribute-specific information. Define the FS-VAE latent code as $\mathbf{z} \in \mathbb{R}^F$ and a scoring vector $\mathbf{s} \in \mathbb{R}^F$, where $0 \leq s_i \leq 1$ for $i \in \{0, \dots, F-1\}$. During FS-VAE training, a weighted feature vector is generated by $\mathbf{z}' = \mathbf{s} * \mathbf{z}$, where $*$ denotes an element-wise product, and then fed into the classifier for attribute-specific information extraction. During the optimization step, the dimensions of the latent code with higher scores for emotion attribute will be updated more when back-propagating emotion classification loss, (hence, containing more emotion-related information), and vice versa. In this work, two FSMs are trained to capture the emotion-related attribute and identity-related attribute, respectively.

Further, to purify the latent dimensions after attribute alignment, we apply Gradient Reversal Layer (GRL) [23]. We restrict the reversed gradients to affect those nodes with significant importance to the particular task. During training, we input

a masked feature vector $\tilde{\mathbf{z}} = \mathbf{m} * \mathbf{z}$ to the discriminator, where the mask vector \mathbf{m} is generated by a threshold function $f(s)$:

$$m_i = f(s_i) = \begin{cases} 1, & s_i \geq \theta_t \\ 0, & s_i < \theta_t \end{cases} \quad (2)$$

Here, θ_t represents the threshold value for the designated task t . Node-specific adversarial learning forces the latent dimension with critical importance to the designated attribute carry ‘‘pure’’ information.

2.3.2. Additional losses

The sensitive attributes, i.e., emotion and identity, can be correlated, the distribution of the scoring vectors from two vanilla FSMs are highly overlapping. Therefore, we design additional constraints to ensure each node is distinctively responsible for a single attribute. We integrate two different losses, attention penalty loss and diversity loss, to guide the attention weights for attribute-alignment. Attention penalty loss is a regularization term that encourages dissimilarity of score vectors between different tasks to prevent redundancy [24]. We define a score matrix $\mathbf{S} \in \mathbb{R}^{|\mathbf{T}| \times F}$, where each row of the matrix is a 12-normalized score vector $\hat{\mathbf{s}}_t$ corresponding to an attribute $t \in \mathbf{T}$. Note that \mathbf{T} is the attribute set with $|\mathbf{T}|$ attributes, where $|\mathbf{T}| = 2$ in this work. Attention penalty loss is defined as:

$$L_{att} = \|(\mathbf{S}\mathbf{S}^T - \mathbf{I})\|_F \quad (3)$$

Here $\|\cdot\|_F$ stands for the Frobenius norm of a matrix and $\mathbf{I} \in \mathbb{R}^{|\mathbf{T}| \times |\mathbf{T}|}$ is an identity matrix. The element s_{ij} in $\mathbf{S}\mathbf{S}^T$ is the cosine similarity between $\hat{\mathbf{s}}_i$ and $\hat{\mathbf{s}}_j$. As the diagonal elements are always equal to 1, we subtract an identity matrix to exclude the diagonal elements in loss calculation. Since non-diagonal elements contribute positive values to the penalty term when the scores are non-orthogonal, through minimization, it encourages dissimilarity between attribute-specific FSM scores.

Diversity loss is another loss used to enhance the intra-class compactness and inter-class separability in the representation space. In this work, we utilize the additive-margin softmax (AM-Softmax) [25], where the classes are the attributes in our case. AM-Softmax imports angular margin into softmax, which forces the model to learn between-class large-margin representations. The diversity loss is derived as:

$$L_{div} = \frac{-1}{N|\mathbf{T}|} \sum_{i=1}^N \sum_{t=1}^{|\mathbf{T}|} \log \frac{e^{s(\hat{\mathbf{W}}_{yt}^T \hat{\mathbf{z}}_i^t - m)}}{e^{s(\hat{\mathbf{W}}_{yt}^T \hat{\mathbf{z}}_i^t - m)} + \sum_{j=1, j \neq yt}^{|\mathbf{T}|} e^{s\hat{\mathbf{W}}_j^T \hat{\mathbf{z}}_i^t}} \quad (4)$$

where N is the number of samples, $\hat{\mathbf{z}}_i^t$ is the weighted feature vector related to attribute t , and $\hat{\mathbf{W}}$ is the last fully connected

Table 1: Privacy-preserving performance presented in WFS (%) and EER (%) for SER and SV respectively, where PP stands for privacy-preserving. Superscript * shows p -value < 0.05 , comparing to the proposed method (FS-VAE) in the last row.

Method	Origin		PP-SER		PP-SV	
	WFS	EER	WFS	EER	WFS	EER
Wav2vec2.0	56.73	12.77	55.81*	21.69*	55.08*	12.96*
A-VAE	-	-	53.22*	35.40*	38.44*	20.70*
D-VAE	53.89	12.50	52.35*	31.21*	38.75*	18.24*
LR-VAE[13]	54.20	14.26	53.25*	34.74*	36.22	17.42*
W-VAE	-	-	52.46*	34.64*	40.72*	17.70*
FS-VAE	54.33	13.68	55.11	32.48	36.52	16.34

layer. Note that scaling factor s and margin m are hyperparameters. This loss ensures that the weighted vectors related to different attributes are separated and condensed on their respective vector spaces. In summary, the complete loss function for FS-VAE is defined as:

$$L_{overall} = L_{VAE} + L_{att} + L_{div} + L_{emo} + L_{id} + L_{emo-adv} + L_{id-adv} \quad (5)$$

where L_{emo} and L_{id} represents the classification loss; and $L_{emo-adv}$ and L_{id-adv} represents the adversarial loss.

2.4. Masking attributes for recognition tasks

In downstream tasks, we first obtain an attribute-aligned representation by the FS-VAE encoder. Then, task-specific scoring machine FSM_t is applied to rank the importance of each node related to task t . With the scores, it is straightforward to perform attribute-selection by masking the dimensions i where $s_i > \theta_t$. For example, to protect identity in PP-SER, we mask the nodes with identity scores greater than θ_{id} .

3. Experiments

3.1. Experiment setup

The FS-VAE model structure is defined as follows: the encoder and decoder are multi-layer perceptrons (MLP) with two fully connected layers modelling the mean and log variance of the latent code. We train two MLP classifiers on emotion recognition and speaker verification tasks, and two MLP discriminators for adversarial GRL learning. We select PReLU as activation function. Moreover, two FSMs are built using MLP. We apply Adam optimizer with learning rate $5e^{-4}$ and batch size 128 for training. Also, weight decay $1e^{-6}$ is added to stabilize the training process. The complete loss function is defined in equation 5. We use mean squared error loss for reconstruction loss, and cross entropy loss for both classification and adversarial loss.

Weighted f-score (WFS) is used to evaluate the performance of SER, and equal error rate (EER) is used to evaluate the performance of SV. We optimize the hyperparameters on the development set and present results on the test set. In this work, we study the following two setups:

PP-SER: Privacy-preserving emotion recognition that aims to protect speaker identity while preserving SER performance (WFS should be high, and EER should also be high).

PP-SV: Privacy-preserving speaker verification that aims to hide the emotion while maintaining the SV performance (EER should be low, and WFS should also be low).

Note that the models for privacy evaluation are trained by masked representation. Also, we test the statistical significance of the difference between different methods' performance. By pairing the proposed method results with different baseline methods and model-variations, we then use Wilcoxon signed-rank test to obtain the p -value for paired difference test. The

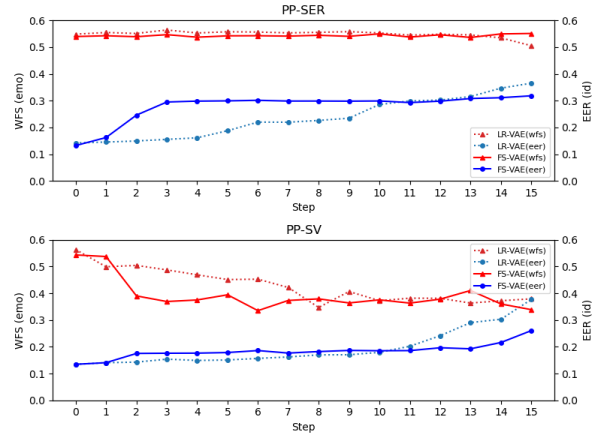


Figure 2: The performance curves in the masking experiment. Note that the y-axis of SER curves are WFS, and EER for SV. For LR-VAE, we mask in bottom-up order in PP-SER, and top-down order in PP-SV. For FS-VAE, we mask from the dimension with higher scores.

results are shown in table 1 and table 2, where the superscript * indicates a statistically significant difference ($p < 0.05$).

3.1.1. Baseline methods

Here, we describe baseline representation learning methods to compare with our proposed FS-VAE:

Wav2vec2.0: Apply original wav2vec2.0 embedding for SER and SV model training. Extract the pre-final layer of SER (SV) model for PP-SER (PP-SV) task.

A-VAE: Apply adversarial learning method [11] using GRL to remove identity (emotion) attributes for PP-SER (PP-SV).

D-VAE: Apply disentangled representation learning, which divides the latent vector into attribute-specific regions. Mask the particular region to achieve privacy protection.

LR-VAE: Proposed method in [13], the most recent SOTA attribute-aligned speech representation learning method.

W-VAE: Apply the weighted latent vector for downstream tasks without attribute-selection (a model variant of FS-VAE).

3.2. Result and analysis

3.2.1. Privacy-preserving performance

Our goal is to protect a particular sensitive attribute while maintaining the main task utility. Comparing to other baselines, our proposed FS-VAE achieves the best performances (PP-SER: 55.11% WFS, 32.48% EER; PP-SV: 16.34% EER, 36.52% WFS) with improved utility and competitive privacy protection. There are a couple observations to note. The first row in table 1 shows that wav2vec2.0 embedding achieves promising results on both tasks of SER and SV, which demonstrates its capability as an informative universal front-end. More interestingly, when we extract the pre-final layer of the SER and SV models to examine the PP-SER and the PP-SV results. The results show that the embedding, even when training for a particular attribute recognition, still contain significant information about other sensitive attributes leading the issue of privacy leakage.

Firstly, we compare the performances to the adversarial representation learning, a prevalent method used for privacy protection. The result is shown in the A-VAE row (table 1). For PP-SER, although FS-VAE achieves a little worse identity protection (an increase of 2.92% EER), it better maintains the SER performance (an increase of 1.89% WFS). On the other hand, for PP-SV, FS-VAE outperforms the A-VAE on both SV performance (a drop of 4.36% EER) and emotion protection (an increase of 1.92% WFS). Note that A-VAE requires scenario-

specific encoders, i.e., retraining an encoder for different protection settings, while FS-VAE requires just a single encoder.

Secondly, we compare our proposed FS-VAE with the SOTA attribute-aligned representation learning method, LR-VAE (the LR-VAE row in table 1). For PP-SER, FS-VAE achieves a competitive result. Although FS-VAE performs slightly worse on identity protection (an increase of 2.26% EER), it results in a better emotion recognition performance (an increase of 1.86% WFS). On the other hand, for PP-SV, FS-VAE shows obvious improved results of better utility maintenance (a drop of 1.08% EER) and slightly worse privacy protection results (a drop of 0.30% WFS). FS-VAE has competitive PP-SER results and improved PP-SV performance when compared to LR-VAE. Note that, FS-VAE is more flexible than LR-VAE in the sense that it can be extended to multiple attribute settings by simply adding additional rows in attention penalty loss and more classes in diversity loss.

Lastly, we compare FS-VAE to disentangled representation learning (D-VAE) and weighted vector without attribute selection (W-VAE). The result is shown in the D-VAE and W-VAE row of table 1 respectively. When comparing to the D-VAE, we observe that FS-VAE achieves better utility maintenance (an increase of 2.76% WFS, a drop of 1.90% EER better) and privacy protection (a drop of 1.27% EER, an increase of 2.23% WFS) for both PP-SER and PP-SV. When comparing to the W-VAE, W-VAE achieves comparable privacy protection as our method though FS-VAE obtains better utility maintenance (an increase of 2.65% WFS, a drop of 1.36% EER). These results highlight the importance of attribute alignment and node masking.

3.2.2. Analysis of attribute-alignment strategy

We conduct a masking experiment to study the effectiveness of feature scoring machine (FSM) for attribute alignment. The usage of the two loss functions guides the attribute-alignment and concentrates the task-specific attributes on particular nodes. We compare the results to the layered dropout strategy [13]. The experiment procedure is as follow: first, we encode input features into latent vectors with 128 dimensions, and sort the dimension by the value of score vector; next, we divide the sorted dimension into 16 groups. During masking, for each step, we mask an additional group of nodes with highest scores; then, the masked latent vectors are applied to two tasks: emotion recognition and speaker verification. For example, in the 1st step, 8 latent dimension with highest scores are masked, while the remaining 120 dimension are applied to both SER and SV models. For LR-VAE, we conduct the same analysis but mask the latent vector from one end to the other end, i.e., the emotion-related end or the identity-related end [13].

We observe the identity-protection emotion recognition task in figure 2, the upper PP-SER row. Comparing the emotion recognition curves (WFS), both LR-VAE and FS-VAE maintain high SER performance as the masking progress moves on, while FS-VAE maintains SER performance better at the 15th step, with only one group left (8 dimension). On the other hand, we can see that the speaker verification curve (EER) of FS-VAE has a rapid increase at the beginning (2nd and 3rd step) of the experiment. It shows that the few top-scored identity-related nodes contains a high portion of speaker identity information with little emotion-related information.

Further, we study the emotion-protection speaker verification task in figure 2, the lower PP-SV row. For the EER curves, the downward trend of FS-VAE is smoother than LR-VAE (utility maintenance). We also observe that LR-VAE experiences an early significant performance drop (12th and 13th step), while

Table 2: Ablation study results. Privacy-preserving performance presented in WFS (%) and EER (%) for SER and SV respectively. Note that ✓ means to include the corresponding component, while – means to exclude the component. Superscript * shows p -value < 0.05 , comparing to proposed method.

Components		PP-SER		PP-SV	
Att-Loss	Div-Loss	WFS	EER	WFS	EER
–	–	40.01*	39.49*	40.47*	37.06*
✓	–	54.15	31.84*	39.28*	16.50
–	✓	52.33*	33.88*	39.99*	22.36*
✓	✓	55.11	32.48	36.52	16.34

FS-VAE has a more gentle downward slope. For WFS curves (privacy protection), we see a rapid performance drop at the first few steps (2nd step) of the experiment indicating that by masking just a few nodes, the emotion-related information in the remaining representation has also effectively been deleted.

3.2.3. Ablation study

We perform ablation study to investigate the effectiveness of the two loss terms, attention penalty and diversity loss. We re-train the FS-VAE with different combinations of these two loss, and apply the latent code for two privacy-preserving scenarios, PP-SER and PP-SV. The results are shown in table 2.

Firstly, we study the baseline case, i.e., training without the two penalty terms. The poor utility maintenance (40.01% WFS, 37.06% EER) shows that without explicit constraints, the two attributes are highly overlapping on similar set of nodes resulting in poor performances. Next, we study the case when applying attention penalty loss only, which is designed to make the attribute-specific scores distinct. The improved utility maintenance (14.14% WFS better, 20.56% EER better) demonstrates that the inclusion of attention penalty loss is key in the alignment. On the other hand, we study the case of applying diversity loss, which enhances the inter-task separability and intra-class compactness in the weighted vectors. The result shows that using diversity loss only loosely encourage the task-specific attributes to distribute on different dimensions, but the constraint is not strong enough as compared to attention penalty loss. Lastly, we observe the case when applying both loss functions where it achieves the best privacy-preserving performance (PP-SER: 55.11% WFS, 32.48% EER; PP-SV: 16.34% EER, 36.52% WFS); the diversity loss, while not enough by itself, can act as an auxiliary term that improve the overall performances.

4. Conclusions

In this work, we propose an attention-based attribute aligned representation learning strategy to achieve flexible speech representation for privacy protection. Comparing to previous methods, it better maintains the utility and achieves competitive performance on PP-SER and improves performance on PP-SV. We also show that the two losses separates task-specific attributes and guides the alignment learning process without explicitly defining dropout functions. These losses enable scoring machines to measure the attribute-specific importance of each dimension and naturally provides a flexible mechanism to select and protect target sensitive attributes. In the future, since our proposed method is extendable to operate in multiple attributes setting, we will immediately evaluate on a proper database with multiple attributes. Moreover, we will generalize our approach from using an aggregated feature vector to time series modeling, and further explore multimodality, such as speech and language, for learning speech representation.

5. References

- [1] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication*, vol. 116, pp. 56–76, 2020.
- [2] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 2014.
- [3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [4] D. Braga, A. M. Madureira, L. Coelho, and R. Ajith, “Automatic detection of parkinson’s disease based on acoustic analysis of speech,” *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 148–158, 2019.
- [5] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákási, and J. Kálmán, “A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech,” *Current Alzheimer Research*, vol. 15, no. 2, pp. 130–138, 2018.
- [6] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, 1st ed. Wiley Publishing, 2013.
- [7] J. L. Kröger, O. H.-M. Lutz, and P. Raschke, *Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference*. Springer International Publishing, 2020, pp. 242–258.
- [8] R. Tatman, “Gender and dialect bias in youtube’s automatic captions,” in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017, pp. 53–59.
- [9] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, “The risk of racial bias in hate speech detection,” in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 1668–1678.
- [10] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, “Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?” in *Proc. Interspeech 2019*, 2019, pp. 3700–3704.
- [11] M. Jaiswal and E. M. Provost, “Privacy enhanced multi-modal neural representations for emotion recognition,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 7985–7993. [Online]. Available: <https://aaai.org/ojs/index.php/AAAI/article/view/6307>
- [12] M. Xia, A. Field, and Y. Tsvetkov, “Demoting racial bias in hate speech detection,” in *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7–14. [Online]. Available: <https://www.aclweb.org/anthology/2020.socialnlp-1.2>
- [13] Y.-L. Huang, B.-H. Su, Y.-W. P. Hong, and C.-C. Lee, “An Attribute-Aligned Strategy for Learning Speech Representation,” in *Proc. Interspeech 2021*, 2021, pp. 1179–1183.
- [14] R. Lotfian and C. Busso, “Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings,” *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October–December 2019.
- [15] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *arXiv preprint arXiv:2006.11477*, 2020.
- [16] L. Pepino, P. Riera, and L. Ferrer, “Emotion Recognition from Speech Using wav2vec 2.0 Embeddings,” in *Proc. Interspeech 2021*, 2021, pp. 3400–3404.
- [17] Z. Fan, M. Li, S. Zhou, and B. Xu, “Exploring wav2vec 2.0 on Speaker Verification and Language Identification,” in *Proc. Interspeech 2021*, 2021, pp. 1509–1513.
- [18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- [19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [20] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [21] N. Gui, D. Ge, and Z. Hu, “Afs: An attention-based mechanism for supervised feature selection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3705–3713.
- [22] B. Škrlj, S. Džeroski, N. Lavrač, and M. Petkovič, “Feature importance estimation with self-attention networks,” *arXiv preprint arXiv:2002.04464*, 2020.
- [23] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.
- [24] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” *arXiv preprint arXiv:1703.03130*, 2017.
- [25] 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.