INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

An Attribute-Aligned Strategy for Learning Speech Representation

Yu-Lin Huang[1,3], *Bo-Hao Su*[1,3], *Y.-W. Peter Hong*[1,2,3], *Chi-Chun Lee*[1,2,3]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]Institute of Communication Engineering, National Tsing Hua University, Taiwan
[3]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

huang8592301@gapp.nthu.edu.tw, borrissu@gapp.nthu.edu.tw,
ywhong@ee.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

Advancement in speech technology has brought convenience to our life. However, the concern is on the rise as speech signal contains multiple personal attributes, which would lead to either sensitive information leakage or bias toward decision. In this work, we propose an attribute-aligned learning strategy to derive speech representation that can flexibly address these issues by attribute-selection mechanism. Specifically, we propose a layered-representation variational autoencoder (LR-VAE), which factorizes speech representation into attribute-sensitive nodes, to derive an identity-free representation for speech emotion recognition (SER), and an emotionless representation for speaker verification (SV). Our proposed method achieves competitive performances on identity-free SER and a better performance on emotionless SV, comparing to the current state-of-the-art method of using adversarial learning applied on a large emotion corpora, the MSP-Podcast. Also, our proposed learning strategy reduces the model and training process needed to achieve multiple privacy-preserving tasks.

Index Terms: speech representation, layered dropout, privacy, fair, attribute alignment

1. Introduction

Speech technology has rapidly proliferated and integrated deeply into our daily life [1, 2, 3, 4, 5]. While these applications bring convenience to our life, several growing concerns have gained attention and need to be addressed with care. The first is *privacy* due to concerns of sensitive information leakage: for example, users may not expect to disclose their identity information while using a speech emotion recognition (SER) system; on the other hand, users may not wish to share their emotional condition when being assessed by a speaker recognition (SR) system. Moreover, the collective social norm would create unwanted and often detrimental self-exaggerated issues around equality, e.g., unfair biases toward gender types [6] or race [7], when using data-driven approaches for speech technology. Speech is an informative signal which contains personal sensitive attributes by nature; hence developing appropriate methods either to protect privacy information, such as identity and emotion, or to mitigate the undesired biases, like gender and race, is critical in the current era.

Recently, several works in speech processing have started to address these issues using privacy-aware representation learning. For example, Srivastava et al. used adversarial representation learning on automatic speech recognition (ASR) to protect speaker identity [8], Alouf et al. used CycleGAN-based method to generate emotion-less synthesized speech for voice assistant to hide personal affect [9], Jaiswal et al. used adversarial learning to generate gender-invariant representation for identify-free emotion recognition [10], and Xia et al. applied adversarial learning to mitigate racial bias in hate speech detection [11]. While current state-of-the-art methods concentrate on using adversarial learning, this strategy suffers from several shortcomings. Adversarial method address privacy issues by learning a speech signal space with no targeted sensitive attributes as measured by its ability in fooling a well-trained discriminator that is in charge of classifying sensitive information, e.g., gender and speaker identity. This attribute invariant learning strategy lacks a flexible mechanism to adapt to different criterion of privacy preserving; for example, in some tasks only the "gender" attribute may need to be protected while some other tasks would require the "speaker identity" to be private. For different scenario of interest, one would have to re-train the adversarial network over again.

In this work, instead of taking a 'per-attribute' adversarial invariant learning approach, we formulate the problem as devising a learning strategy that would result in attribute-aligned speech representation. The core idea centers on conceptualizing that speech contains a mixture of attributes, [12, 13], e.g., gender, age, emotion and semantics, etc. By factorizing the entangled information of speech signal into independent attributes with proper attribute-alignment, we can protect particular sensitive information by attribute selection, i.e., masking targeted sensitive attributes, to minimize either privacy-leakage or biased decision. In this paper, we evaluate this idea by targeting two sensitive attributes in speech, i.e., emotion and identity, and our aim is to show that this approach can flexibly achieve privacy-preserving applications by eliminating identity contents in SER or emotion contents in SV at ease.

We propose a framework of flexible attribute masking for speech, inspired by the fair representation learning [14]. We aim to learn a layered disentangled speech representation with a backbone of variational autoencoder (VAE) [15, 16]. We specifically propose a layered dropout strategy in a multi-task framework to achieve attribute-alignment, i.e., forces the latent to align in an emotion-related to identity-related order. To further *clean up* the aligned representation knowing that these two attributes are highly correlated [17, 18], we add adversarial reversal layer to each task-specific branch. Our strategy provides flexibility in either identity masking or emotion masking to come up with an identity-free latent for privacy-preserving SER or emotionless latent for privacy-preserving SV with a unified learning framework. In this work, we evaluate our method on MSP-Podcast [19] for SER and SV tasks using three types of feature, and achieve competitive results on SER (emobase: 52.41% weighted f-score, 41.14% EER), and an improvement on SV (netvlad: 34.35% weighted f-score, 10.91% EER; x-vector: 34.23% weighted f-score, 9.63% EER), compared to the state-of-the-art adversarial learning method.

http://dx.doi.org/10.21437/Interspeech.2021-1341

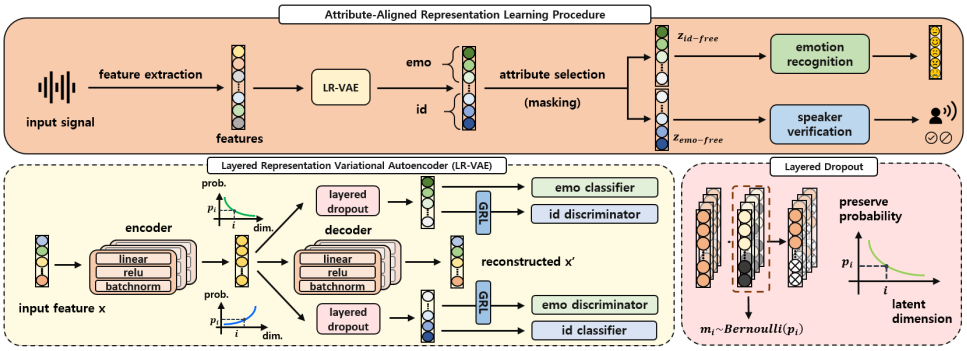


Figure 1: *An illustration of our proposed method for attribute-aligned representation learning. It includes three blocks: representation learning procedure, LR-VAE, and layered dropout. Notice that $Z_{id-free}$ stands for identity-free representation, $Z_{emo-free}$ stands for emotionless representation.*

2. Methodology

2.1. Dataset Description

In this study, we focus on two main tasks, emotion recognition and speaker verification. To evaluate the performance of these two tasks, a large corpus with emotional labels and multiple speakers is needed. Hence, we use the MSP-Podcast database [19], which includes over 1,000 podcast recordings. Each podcast is segmented into speaking turns, where segments with music, overlapped speech, telephone quality speech and background noise are discarded.

In this work, we use data with 5 categorical emotions: neutral, angry, sad, happy and disgust as in [18]. We used the standard splits in Release 1.4 for training, development, and testing, which includes 610 speakers in train set, 30 speakers in development set, and 50 speakers in test set, where each set of speakers are disjoint. The distribution of the 5 emotion classes are: angry: 8.81%, happiness: 27.10%, neutral: 53.05%, sad: 3.95%, disgust: 7.09%.

2.2. Feature Extraction

In this work, we use three different input features for the two tasks: emobase2010, netvlad embedding, and x-vector embedding to verify the effectiveness of our proposed method. First, we use emobase2010, which is a commonly used feature for SER, as input. It is a 1582 dimensional feature including pitch, loudness, mfcc and spectral, etc. We extract emobase2010 using openSMILE toolkit [2]. Further, we extract embeddings commonly used in state-of-the-art speaker verification task, i.e., netvlad [3] and x-vector [18]. The netvlad embedding is extracted using the released pre-trained model [3], while the x-vector embedding is obtained by training on the Voxceleb2 [20] using the structure mentioned in [18].

2.3. Layered Representation Variational Autoencoder

We propose a layered-representation variational autoencoder (LR-VAE) to factorize the entangled dimensions contained in speech and arrange these dimensions in an emotion-related to identity-related order. LR-VAE contains two main components, i.e., disentangled representation and layered dropout. We will first describe VAE, i.e., a well-known structure for disentangled learning. Then, we will further detail our layered dropout with adversarial multitask learning to obtain attribute-aligned speech representation.

2.3.1. Variational Autoencoder (VAE)

In this work, we use disentangled representation learning via VAE to derive a latent node-wise independent representation. VAE model aims to learn the marginal likelihood of a data $\mathbf{x}$, with the objective function:

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

where $D_{KL}(||)$ stands for the non-negative Kullback-Leibler divergence. The KL-divergence term encourages the posterior distribution to be close to an isotropic Gaussian to achieve disentanglement purpose.

2.3.2. Layered Dropout with Adversarial Multitask Learning

In this work, we propose a strategy of layered dropout with a multitask-learning architecture. Multitask learning aims to include both emotion and speaker identity information into the latent codes. Layered dropout is further utilized to force these attributes to align toward both ends of the latent codes resulting in a layered representation. Also, adversarial branches, i.e., gradient reversal layer, are used to additionally 'purify' this attribute-aligned representation.

Dropout is a well-known regularization method in deep learning to prevent neural networks from overfitting[21]. Layered dropout works in a similar manner but with a different purpose. We propose to use this as a learning mechanism to make each dimension of the latent codes carry different importance to the designated task. In our work, the two tasks are defined as the emotion recognition and the speaker verification. We design a dropout rate function making the probability of dropping decreases (or increases) monotonically for each node of the input layer. This effectively forces the target task's discriminatory information to concentrate on nodes with lower dropout rates.

Let $\mathbf{x}$ denotes the input vector with N dimensions of a layer of a neural network, we define a vector with decreasing preserving rates $\mathbf{p}$ for task of emotion recognition (increasing preserving rates for speaker verification), where $0 \leq p_i \leq 1$ for $i \in \{0, \ldots, N-1\}$. With layered dropout, the input vector $\mathbf{x}$ of the feed-forward operation is replaced by vector $\widetilde{\mathbf{x}}$, generated by:

$$m_i \sim \text{Bernoulli}(p_i) \quad (2)$$

$$\widetilde{\mathbf{x}} = \mathbf{m} * \mathbf{x} \quad (3)$$

Here, $*$ denotes an element-wise product. $\mathbf{m}$ acts as a mask before the vector $\mathbf{x}$ is fed into the layer. For a dimension m_i in the vector $\mathbf{m}$, it's an independent Bernoulli random variable with probability p_i being 1, which means to preserve the i^{th} node, and 0 means to drop the i^{th} node. While testing, $\mathbf{W}$

Table 1: *The experiment results are presented in weighted f-scores (%) and EER (%) for SER and SV respectively. Notice that PP stands for privacy-preserving, where columns of origin stand for original representation without privacy protection, PP-SER stands for identity-free SER, and PP-SV stands for emotionless SV.*

Method		DNN	VAE	A-VAE		LR-VAE (w/o adv)			LR-VAE		
		origin	origin	PP-SER	PP-SV	origin	PP-SER	PP-SV	origin	PP-SER	PP-SV
emobase	emo	54.90	52.61	52.09	37.93	53.54	**52.71**	40.33	52.86	52.41	**37.54**
	id	14.45	12.99	**41.49**	17.16	13.27	30.64	19.16	11.77	41.14	**12.70**
netvlad	emo	52.99	49.80	**49.24**	38.92	50.11	49.04	**34.10**	50.01	48.23	34.35
	id	8.14	8.37	40.66	15.25	8.20	32.39	14.51	8.53	**49.51**	**10.91**
x-vector	emo	53.24	52.60	48.95	**34.13**	52.22	**51.76**	36.05	52.57	51.09	34.23
	id	10.05	8.55	45.93	13.56	8.75	37.61	10.66	8.44	**49.48**	**9.63**

represents for weights of network and the weights are scaled as $\mathbf{W}_{test} = \mathbf{W} * \mathbf{p}^{\top}$ and inference without dropout, which is same as the vanilla dropout layer.

This layered dropout mechanism alters the dropout rates being applied on both sides of the representation before an emotion (identity) classifier, the latent codes form an aligned emotion-to-identity order from top end to bottom end during the optimization step. Furthermore, we add an auxiliary mechanism of adversarial branches with gradient reversal layers [22] during multitask learning. The goal is to learn cleaner factorized identity-free (emotion-free) representation. After having an attribute-aligned representation, we simply need to mask the dimension representing the particular attribute of interest. For example, to protect identity information in SER, we can simply mask the nodes that have high emotion preservation rates and low speaker identity preservation rates, and vice versa for in SV. Notice that our attribute aligned strategy provides a mechanism to select "what to protect" with a single unified learning, which is more efficient than the adversarial method that requires re-training in different scenarios.

3. Experiments

3.1. Experiment Setup

The structure of our VAE model is as follows: multi-layer perceptron (MLP) is applied for encoder and decoder. Additionally, fully connected layer is applied to model the mean and log variance of the latent code for the encoder. For multi-task learning, two MLP classifiers are trained for emotion recognition and speaker identification, and two MLP discriminators are trained for adversarial learning by applying gradient reversal layers [22]. We set the learning rate as $5e^{-4}$, and the batch size as 128. Moreover, we add a regularization of $1e^{-6}$ to all weights and biases to stabilize the training process. Rectified Linear Unit (ReLU) is chosen as the activation function. We train the model using Adam optimizer, with L_{obj} as the objective, defined as:

$$L_{obj} = L_{VAE} + L_{emo} + L_{id} + L_{emo-adv} + L_{id-adv} \quad (4)$$

where L_{VAE} represents the reconstruction error and KL divergence loss as defined in equation 1, while L_{emo} and L_{id} represents the cross-entropy loss for emotion recognition and speaker identification; $L_{emo-adv}$ and L_{id-adv} represents the adversarial loss for emotion recognition and speaker identification. Notice that for speaker verification (SV) task, models are trained to predict speaker identity in the training set, i.e., speaker identification, to learn identity-related information during training; while during evaluation, the hidden layer embedding is extracted and apply to speaker verification system.

We evaluate the performance of SER using weighted f-score (WFS), following the experiment setup in [18], and evaluate the performance of speaker verification by equal error rate (EER). For each feature set, we train a factorized layered representation encoder based on training set, select model using validation set, and test performance on test set. Assuming attackers have access to the training set with encoded representation and labels of speakers. Our goal is to generate a representation such that for the encoded representations with particular sensitive attributes masked, neither attackers nor hosts are able to identify the sensitive attributes while the main task performance is maintained.

3.1.1. Baseline Methods

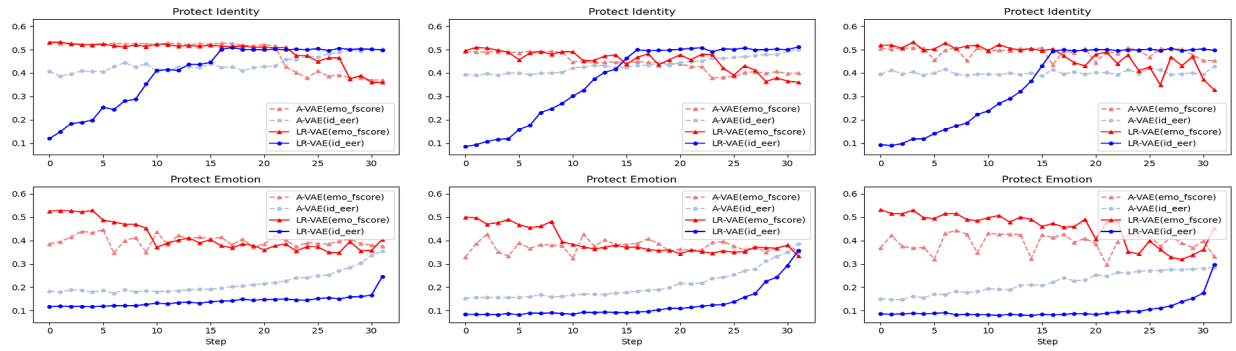
The following are the baseline methods of different learning strategies that we use to compare with LR-VAE. Notice that privacy-preserving (PP) on LR-VAE are done by masking the dimension of particular sensitive attributes in the latent codes.
DNN: A model conducted by fully connected layers to obtain the baseline performance on SER and SV for each feature.
VAE: A vanilla VAE trained by multi-task learning on SER and SV tasks.
A-VAE: A VAE trained for single task (SER or SV) with adversarial learning (reverse gradient) on the other task (SV or SER).
LR-VAE (w/o adv): A model similar to our proposed LR-VAE, but trained without adding the adversarial branch.

3.2. Result and Analysis

3.2.1. Sensitive Attribute Protection

Note that all the comparison are presented in absolute points in this section. For privacy-preserving speech emotion recognition (PP-SER), we aim to protect user's identity information while preserving the emotion recognition performance. As shown in the PP-SER columns in table 1, our proposed LR-VAE achieves the better privacy preserving performance on x-vector and netvlad and similar result on emobase comparing to A-VAE. It shows that our proposed method is able to obtain a competitive emotion recognition performance (0.32% WFS higher on emobase, 1.01% WFS lower on netvlad, and 2.14% WFS higher on x-vector), with better improvements in protecting speaker identity (only 0.35% EER worse on emobase, 8.85% EER better on netvlad, and 3.55% EER better on x-vector).

On the other hand, to achieve emotion-protected speaker verification (PP-SV), we aim to reduce users' emotional information in the speech while preserving the speaker verification performance. As shown in PP-SV columns in table 1, our proposed LR-VAE achieves the best emotion protection performance on all three features comparing to A-VAE. It shows that our proposed method could better maintain the speaker verifica-



(a) *emobase* (b) *netvlad* (c) *x-vector*

Figure 2: *The performance curves in the masking experiment where the y-axis for SER results are weighted f-scores, and EER for SV. For the first row (protect identity), the masking process starts in a bottom up order, while for the second row (protect emotion), we start masking from the top group.*

tion performance (4.46% EER better on emobase, 4.34% EER better on netvlad, and 3.93% EER better on x-vector), while achieving state-of-the-art emotion-related attributes protection (0.39% WFS better on emobase, 4.57% WFS better on netvlad, and 0.10% WFS worse on x-vector)

We first study the baseline DNN results shown in the column, DNN, in table 1. The promising performance show that regardless of features, it contains both emotion and identity information. It reinforces the current concerns that speech contains many personal attributes that users may not want to reveal. Then, we compare the DNN results to VAE results shown in the column, VAE origin, in table 1. We do see that there is a slight performance drop in emotion recognition potentially due to the information loss caused by kl-divergence loss in VAE training for factorization, which is a trade-off between disentanglement and reconstruction. This factorization VAE is however a key backbone in achieving our attribute-aligned representation.

To study how adversarial branches work in our framework, we compare LR-VAE results to LR-VAE(w/o adv). It shows that without adversarial learning in explicitly purifying the emotion-related (identity-related) dimension to identity-free (emotion-free), the representation learned is not "clean" enough. Hence, while LR-VAE(w/o adv) also achieves competitive results on main tasks, the sensitive attribute-preserving results are usually worse. This also demonstrates that the emotion-related (identity-related) attributes may contains identity (emotion) information if not explicitly cleaned.

3.2.2. Analysis of Aligned Attributes

In this section, we further discuss the effectiveness of layered dropout that align attribute-specific information to both ends of the latent codes. We conduct an experiment with the following procedure: we encode the input features into latent codes; next, we divide the latent dimensions into 32 groups; then, for each step, we mask one additional group of latent codes, and train two models, one for emotion recognition, and the other for speaker verification. We compare the performance curve of LR-VAE and A-VAE to observe how layered dropout influence the discriminatory power of the chosen latent code dimension.

We first study the privacy-preserving speech emotion recognition task. The results are shown in the upper row of figure 2. In this experiment, we start masking from the bottom of the latent code, which contains more identity-related attributes, to the top of the latent code (more emotion-related attributes). As the procedure moves on, the speaker verification performance steadily decreases (EER increases) until the masking process reaches the middle part of the latent code, where it results in a high EER indicating the point where we achieve an identity-free representation. We can also see that EER curves of LR-VAE and A-VAE intersects, which shows that the masked LR-VAE latent can better eliminate the identity-related attributes. On the other hand, we also observe that the emotion recognition performance slightly decreases toward the ending portion of masking process due to a significant reduction in the node dimension, though A-VAE has an even earlier performance drop.

Next, we study the emotion-protection speaker verification task. The results are shown in the lower row of figure 2. In this part of experiment, we start masking from the top of the latent code, similar to the previous procedure, but in a reverse order. As the progress moves on, the emotion recognition performance steadily decreases (weighted f-score decreases), and finally reaches to a similar result comparing to A-VAE. On the other hand, we can see that the speaker verification performance of LR-VAE is better preserved comparing to A-VAE, i.e., the EER curve of LR-VAE is lower in the beginning and increases slower comparing to A-VAE.

4. Conclusions and Future works

In this paper, we propose a novel disentangled layered speech representation learning that can flexibly preserve sensitive attribute in a unified single training architecture. Compared with other methods, our method achieves a competitive performance on identity-free SER and an improvement on emotionless SV. Also, we show that our proposed method help in pushing the emotion and identity information toward the both ends of the latent codes, and this strategy provides a flexible mechanism to select the target sensitive attributes to protect. Moreover, our attribute aligned learning strategy reduce the training and memory cost as we require only single process and single model to achieve competitive privacy-preserving results on SER and SV against adversarial training, which requires training twice and two models.

In the future, we will generalize our attribute aligned representation from two specific task to general multi-attributes scenarios. We could utilize the middle portion of the latent codes to capture other information about the speaker, e.g., gender, personality, semantics, etc, in order to provide a more complete profile on this factorized speech representation. Moreover, as the disentanglement achieved by kl divergence loss causes information loss, different factorization methods may be applied to enhance our representation capacity.

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] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in Proceedings of the 18th ACM International Conference on Multimedia, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 14591462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[3] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5791–5795.

[4] D. Braga, A. M. Madureira, L. Coelho, and R. Ajith, "Automatic detection of parkinsons 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. Pákáski, 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] R. Tatman, "Gender and dialect bias in youtubes automatic captions," in Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, 2017, pp. 53–59.

[7] 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.

[8] 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. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2415

[9] R. Aloufi, H. Haddadi, and D. Boyle, "Emotionless: privacy-preserving speech analysis for voice assistants," arXiv preprint arXiv:1908.03632, 2019.

[10] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in AAAI, 2020.

[11] 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

[12] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in Proc. Interspeech 2017, 2017, pp. 1273–1277. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-349

[13] L. Li, D. Wang, Y. Chen, Y. Shi, Z. Tang, and T. F. Zheng, "Deep factorization for speech signal," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5094–5098.

[14] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," in Proceedings of the 36th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 1436–1445.

[15] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.

[16] ——, "An introduction to variational autoencoders," Found. Trends Mach. Learn., vol. 12, no. 4, pp. 307–392, 2019. [Online]. Available: https://doi.org/10.1561/2200000056

[17] M. Bancroft, R. Lotfian, J. Hansen, and C. Busso, "Exploring the intersection between speaker verification and emotion recognition," in 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2019, pp. 337–342.

[18] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7169–7173.

[19] 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, 2019.

[20] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in Proc. Interspeech 2018, 2018, pp. 1086–1090. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1929

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res., vol. 15, no. 1, p. 19291958, Jan. 2014.

[22] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in International conference on machine learning. PMLR, 2015, pp. 1180–1189.