Interspeech 2022
18-22 September 2022, Incheon, Korea

Investigating the contribution of speaker attributes to speaker separability using disentangled speaker representations

Chau Luu, Steve Renals, Peter Bell

Centre for Speech Technology Research, University of Edinburgh, UK
{chau.luu, s.renals, peter.bell}@ed.ac.uk

Abstract

Deep speaker embeddings have been shown to encode a wide variety of attributes relating to a speaker. The aim of this work is to separate out some of these attributes in the embedding space, disentangling these sources of speaker variation into subsets of the embedding dimensions. This is achieved modifying the training procedure of a typical speaker embedding network, which is typically only trained to classify speakers. This work instead adds pairs of attribute specific task heads to operate on complementary subsets of the speaker embedding dimensions. While specific dimensions are encouraged to encode an attribute, for example gender, the other dimensions are penalized for containing this information using an adversarial loss. We show that this method is effective in factorizing out multiple attributes in the embedding space, successfully disentangling gender, nationality and age. Using the disentangled representations, we investigate how much removing this information impacts speaker verification and diarization performance, showing that gender is a significant source of separation in the deep speaker embedding space, with nationality and age also contributing to a lesser degree.

1. Introduction

Speaker embeddings are a crucial component in many speaker recognition pipelines, with extracting speaker discriminative features being a key step in speaker verification and diarization. In recent years, obtaining speaker embeddings from the intermediate layer of a neural network (x-vectors) has become the leading method for both tasks [1, 2], outperforming the traditionally successful i-vector technique [3].

There are many properties of speech that convey a speaker's identity, including factors related to the physical properties of the vocal apparatus producing the speech (influenced by factors such as gender, age or medical conditions), in addition to properties relating to accent, dialect, native language and sociolect (social or professional group, which can determine lexicon, syntax, stylistics). Humans, upon hearing a new voice, can intuitively infer many of these properties. It is these properties that are used to distinguish between speakers when speaker classification is performed by human experts for criminal cases, a field of practice known as forensic phonetics and acoustics [4, 5].

What this work aims to explore is whether these speaker attributes can be disentangled in the embedding space, and if so, also to determine the contribution that each attribute has on speaker separability. The definition of disentangled representations can be somewhat unclear, but generally speaking, disentangled representation learning aims to learn representations that axis aligns with the underlying generative factors of the data [6, 7, 8]. The exact criteria that determine what constitutes 'generative factors of the data' is under debate [9], but in the context of speaker representations and the human voice, we suggest the factors supported by forensic phonetics literature, like gender, age and accent, are excellent candidates for generative factors that constitute speaker identity. To be explicit, this would mean specific dimensions of the speaker embedding would describe these generative factors in their entirety.

In order to achieve disentangled speaker embeddings in a supervised fashion, this work proposes an architecture that adds pairs of attribute specific task heads alongside the standard speaker classification objective to the standard speaker embedding network. Each pair consists of a predictor and an adversary, which act on complementary dimensions of the embedding, simultaneously encoding attribute information in the chosen dimensions while also removing it from the remaining dimensions.

Using these disentangled embeddings, this work also seeks to understand how information about the gender, age or nationality of a speaker contributes towards the discriminative performance of embeddings in verification and diarization applications. This is explored by evaluating on the VoxCeleb [10, 11] dataset, along with US Supreme Court recordings.

2. Related Work

In the work of [12], speaker representations were disentangled into style and speaker factors using an dual pathway auto-encoder architecture, which used multi-task learning to encourage two auto-encoder latent spaces to separate out these two factors. This work looks at speaker embeddings with a similar approach, but focuses on speaker-specific sources of variation, in addition to incorporating adversarial training techniques to ensure disentanglement.

Both deep speaker embeddings and i-vectors have already been shown to encode a wide variety of information and meta-information about speakers and utterances, such as speaking style and emotion [12, 13], accent and language [14] or speaker gender, channel and transcription information [15]. Furthermore, in [16], it was shown that explicitly encouraging the speaker embedding space to capture nationality and age using multi-task learning could lead to more robust performance on unseen speakers. While [16] looked at improving embedding performance by adding auxiliary speaker-attribute tasks, this work looks to disentangle and probe these attributes by using similar techniques.

The topic of disentangled speaker representations is also closely linked with the field of voice privacy [17, 18, 19], wherein certain attributes are desirable to obscure in speaker embeddings to protect against malicious attackers. Notably, the work of [20] used adversarial training to control the gender element of an auto-encoder architecture, seeking to be able to control that element and therefore provide gender-invariant representations. A follow up paper [21] utilized normalizing flows to again obscure the gender information in speaker embeddings,

10.21437/Interspeech.2022-10643

finding this to be an improvement over the adversarial method.

3. Methodology

3.1. Multi-task Learning

Multi-task learning (MTL) is a learning paradigm in which the same representation can be used to solve multiple different tasks on the same data. For deep learning, this typically means the initial layers of a neural network are shared between tasks, after which task specific layers act on the same shared intermediate representation of the input data.

In the context of training speaker embeddings, such as for the x-vector network [1, 2], these embeddings are extracted from the intermediate layer of a network trained on speaker classification. If the layers up until the embedding are viewed as the embedding extractor, one can consider the remaining layers to be a task specific 'head'. The unmodified x-vector network has a single task specific head, a feed forward network performing speaker classification. MTL can be applied by adding separate task-specific heads with their own loss functions which also act on the embedding.

An MTL speaker embedding architecture can be trained as a whole by optimizing based on a weighted sum of each loss produced by the task-specific heads. If we consider M additional tasks, the multi task loss can be viewed as follows:

$$\mathcal{L}_{\text{multi-task}} = \mathcal{L}_{\text{speaker}} + \sum_{i=1}^{M} \lambda_m \mathcal{L}_{\text{m}} , \tag{1}$$

where each additional task loss $\mathcal{L}_{\text{m}}$ is weighted by some chosen loss weighting λ_m, relative to the speaker classification loss $\mathcal{L}_{\text{speaker}}$.

3.2. Adversarial Training

Domain adversarial training [22, 23] is a technique that involves training an adversary to ascertain the domain of the generated features, such as embeddings. To train a speaker classification network in this fashion, a feed forward network would be used as the adversary, using the embedding layer as an input, and performing a task such as domain classification (Figure 1b). This task would produce an adversarial loss, and would be added to the overall loss function like so:

$$\mathcal{L}_{\text{Domain-Adversarial}} = \mathcal{L}_{\text{Speaker}} - \lambda^{\text{adv}}_{\text{Domain}} \mathcal{L}^{\text{adv}}_{\text{Domain}} , \tag{2}$$

where λ^{adv} is a controllable parameter to determine the weighting of this loss term. Allowing the adversary to act against the rest of the network is implemented via a gradient reversal layer (GRL) between the extractor and the discriminator, which multiplies the gradients by a negative constant during back-propagation. As the loss weighting is negative, the overall loss is reduced if the adversary cannot perform the domain task successfully (high $\mathcal{L}^{\text{adv}}_{\text{Domain}}$). The result is an embedding space which penalizes the inclusion of domain specific information, in theory increasing robustness to changes of domain.

From this formulation, it is easy to see how adversarial training can be viewed in conjunction with MTL. In this regard, an adversary can be considered to be a kind of task-specific head, with a negative loss weighting and a gradient reversal layer.

3.3. Disentanglement

This work proposes a means of utilizing both MTL and adversarial techniques to encourage the speaker embedding space to factorize out specific sources of speaker variation. This is achieved by having auxiliary task heads act on subsets of the full speaker embedding dimensions, supplemental to the standard speaker classification head which takes in the full embedding as input. In this system, each factorized speaker attribute would have a pair of task heads, a predictor and an adversary with a gradient reversal layer.

For example for gender, if we would like to factorize out this attribute into the first dimension of the speaker embedding, the first dimension would be used as input to the predictor, a standard classification head that predicts the gender of the speaker. Simultaneously, the remaining dimensions of the embedding would be input into the adversary that is also predicting gender. By doing so, the first dimension is encouraged to be predictive of gender, while the rest of the speaker embedding is penalized for containing this information - thereby factorising out this speaker attribute.

Importantly, all dimensions are still used as input to the speaker classification head, meaning all sources of variation can be used in performing speaker classification. Figure 1c displays how the proposed system could be trained to factor out Gender and Age into the first and second dimensions of a speaker embedding respectively[1].

4. Experimental Setup

The two datsets used in this work were VoxCeleb [10, 11] and the Supreme Court of the United States (SCOTUS) oral arguments corpus [24], which have web-scrapable speaker attribute information about nationality and age respectively (and both having gender labeling). More information on SCOTUS can be found in [16].

The architecture chosen for the speaker embedding extractor was the x-vector architecture, which was trained on VoxCeleb 2 for 200,000 iterations. The number of embedding dimensions was chosen at 64. For all experiments in which the embedding dimension was split up (referred to as SplitDim), the first embedding dimension was always used to capture the Gender. For VoxCeleb SplitDim experiments, dimensions 2-12 were used as input for the nationality classification task, and for SCOTUS SplitDim experiments, dims 2-12 were re-purposed for a 10-bin age classification task.

SplitDim experiments were also performed without the addition of the adversaries, denoted by Adv or No-Adv. A baseline was also trained which only had a speaker classification head (Figure 1a). Models evaluated on SCOTUS were fine-tuned on SCOTUS from the VoxCeleb model for 20,000 iterations. The following values were chosen for each loss weighting: $\lambda_{\text{Gender}} = 0.05$, $\lambda^{\text{adv}}_{\text{Gender}} = -20.0$, $\lambda_{\text{Nationality, Age}} = 0.05$, $\lambda^{\text{adv}}_{\text{Nationality, Age}} = -10.0$.

To establish the effectiveness of the proposed method of disentangling speaker attributes, both qualitative and quantitative approaches were taken. Firstly, the embedding spaces were examined using t-SNE [25], varying which dimensions to include in this visualization, and labeling points based on supposedly disentangled attributes. Furthermore, the embeddings were probed for information by training a separate feed forward neural network on 50,000 embeddings (as fixed inputs) from the

[1] `https://www.github.com/cvqluu/splitdim_disentangle`

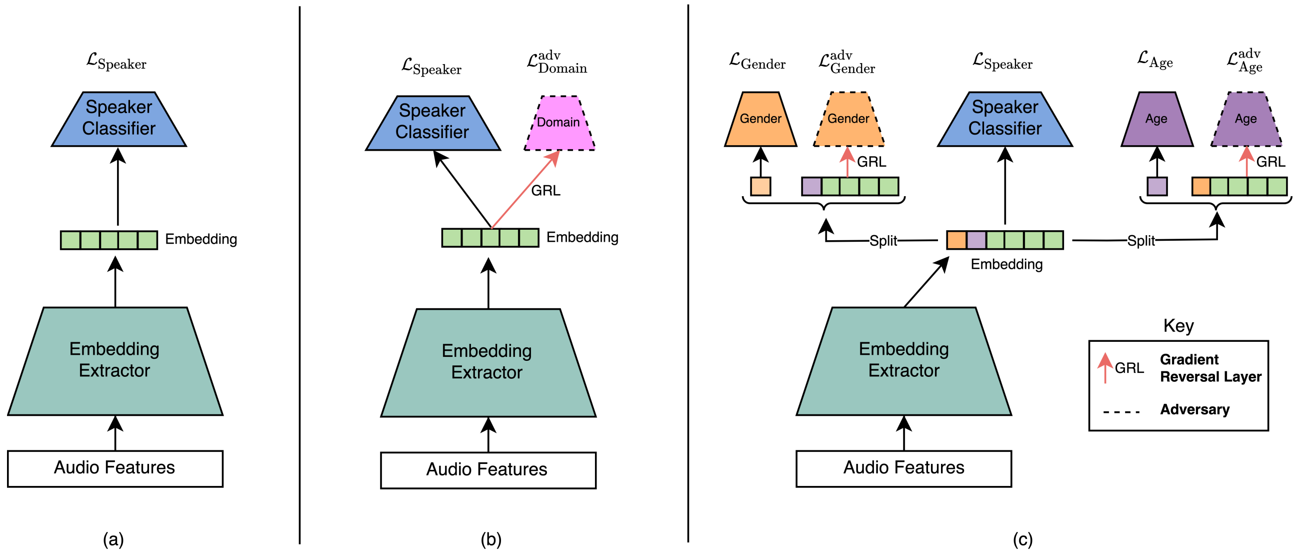

$\mathcal{L}_{\text{Speaker}}$
Speaker Classifier
Embedding
Embedding Extractor
Audio Features
(a)
$\mathcal{L}_{\text{Speaker}}$
$\mathcal{L}^{\text{adv}}_{\text{Domain}}$
Speaker Classifier
Domain
GRL
Embedding
Embedding Extractor
Audio Features
(b)
$\mathcal{L}_{\text{Gender}}$
$\mathcal{L}^{\text{adv}}_{\text{Gender}}$
$\mathcal{L}_{\text{Speaker}}$
$\mathcal{L}_{\text{Age}}$
$\mathcal{L}^{\text{adv}}_{\text{Age}}$
Gender
Gender
GRL
Speaker Classifier
Age
Age
GRL
Split
Split
Embedding
Embedding Extractor
Audio Features
Key
GRL Gradient Reversal Layer
Adversary
(c)

Figure 1: *The architecture for training a: (a) standard speaker embedding extractor, (b) domain adversarial speaker embeddings, (c) speaker-attribute disentangled speaker embeddings (SplitDim-Adv).*

training set, and then evaluating on the test set. If for example a separate classifier was able to perform gender classification successfully on the non-gender dimensions of the embedding, it would imply that the disentanglement had not been successful, as this information remained in the other dimensions. Similarly, the opposite observation would demonstrate that gender information was successfully removed and factored out into the desired dimension.

After showing a suitable level of disentanglement, the verification and diarization performance of these models were evaluated in terms of Equal Error Rate (EER) and Diarization Error Rate (DER). This was evaluated while removing certain attributes (dimensions) from the embeddings, and thus demonstrating what each attribute might contribute to the overall speaker separability. To account for the performance change from removing dimensions of the embedding alone, dimensions were removed from the baseline model to find the average new performance with a reduced number of dimensions, sampling at maximum 1000 permutations of the dropped dimensions. All embeddings were normalized and scored using cosine similarity. Diarization was performed using agglomerative hierarchical clustering with linkage threshold tuned on train-set recordings, extracting embeddings for 1.5s windows with 0.75s stride from oracle speaker activity boundaries.

5. Results and discussions

The t-SNE plots of the embeddings produced by various models can be seen in Figure 2. Here, one can see that in almost all embedding spaces, the separation of embeddings by gender is clearly visible, and this includes Figure 2b, showing that the SplitDim without adversaries still encodes gender in the remaining embedding dimensions. SplitDim-Adv however (Figure 2c), improves in this regard, as when removing the gender dimension, shows much less clear separation between embeddings from each gender. This indicates the necessity of including the adversary to ensure such an attribute is truly disentangled in the embedding.

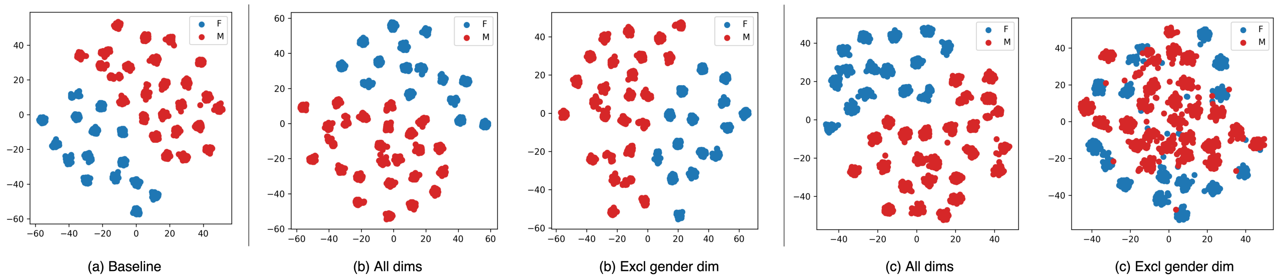
This idea is confirmed further with Table 1, in which a separate classifier was used to probe the embeddings for gender and nationality information. Here, the probing classifier was unable to achieve high accuracy on gender classification when the gender dimension was removed from the SplitDim-Adv embeddings, reducing the probed gender accuracy from 99.47% to 67.08%, which is less accurate than always predicting the most probable test set gender (70.77%).

		Probed Accuracy	
		Gender	Nat.
	Baseline	99.47%	75.92%
	Pick most probable class	70.77%	59.55%
No-Adv	All dims	99.36%	73.38%
	-Gender dim	98.27%	73.01%
	-Nationality dims	98.35%	70.39%
	-Nationality, Gender dims	98.48%	68.52%
Adv	All dims	99.49%	72.38%
	-Gender dim	67.08%	72.86%
	-Nationality dims	97.31%	60.04%
	-Nationality, Gender dims	64.18%	58.38%

Table 1: *The gender and nationality accuracies on VoxCeleb when training a separate probe classifier on embedding features, removing dimensions.*

For SplitDim-no-Adv, gender accuracy was still very high, even when removing the gender dimension, again suggesting that without the adversary, gender information is still present in the other embedding dimensions. These conclusions also carry over to the results with probed nationality accuracy, where the addition of the adversary (Adv) resulted in a much more significant reduction in probed accuracy when removing the relevant dimensions, compared to not (No-Adv).

On a practical note, it should be mentioned that attempts to add the task specific heads to an embedding extractor pretrained on only speaker classification were unsuccessful, resulting in embedding spaces that could not be disentangled. SplitDim-Adv models were only successful when training from scratch or by initializing using another SplitDim-Adv model (as was the case when fine-tuning SplitDim-Adv from VoxCeleb to SCOTUS). This could explain the findings of [21, 20],



(a) Baseline

(b) All dims

(b) Excl gender dim

(c) All dims

(c) Excl gender dim

Figure 2: *Gender color-coded t-SNE projections of the embeddings produced by: (a) baseline model, (b) SplitDim-no-Adv model, (c) SplitDim-Adv model*

which found adversarial techniques to be ineffective in obscuring the gender information when using pre-trained speaker embeddings.

In Table 2, the speaker verification performance on VoxCeleb is shown for the baseline model, along with the SplitDim-Adv model. Firstly, we can see that disentangling the space has incurred a reduction in performance (4.22% to 6.68% EER), which is likely due to the addition of the four extra tasks of the SplitDim-Adv model (Gender, Gender-Adversary, Nationality, Nationality-Adversary). With extra tasks, especially adversarial ones, reaching the optimal embedding space for speaker recognition may be difficult if tasks can conflict with each other (as they are designed to do in adversarial training).

This conflicting performance may also raise questions as to what degree it is possible to fully disentangle certain attributes. For example with age and gender, male and female voices may age in significantly different ways, and thus in order to capture that effectively, the dimensions reserved for predicting age may benefit from containing information about the gender also. This kind of query is very much an open question in disentangled representation learning literature [9], and out of the scope of this paper.

Table 2 also shows the verification performance when removing these attribute specific dimensions. As mentioned in section 4, there is a general performance impact to be expected from removing dimensions in general, and thus the same number of dimensions was also removed from the baseline for a fairer comparison with the removal of attribute specific dimensions. When comparing like for like, the removal of the single gender dimension is significant in comparison to removing a single dimension (1.8% versus 14.2% relative increase in EER), suggesting gender is a powerful contributor to speaker separability, at least in this test set. Likewise, removing Nationality and Nationality with Gender dimensions results in performance degradation beyond that of the baseline model, further supporting that these attributes are significant sources of speaker variation in the speaker embedding space.

For SCOTUS in Table 3, verification performance follows a similar trend to VoxCeleb, with gender once again being a significant factor in affecting separability, whereas the affect that removing age had on performance was more than the baseline expectation from removing 10 dims, but not as significant as nationality. However, for diarization, results are somewhat unexpected, with the SplitDim-Adv model outperforming the baseline in all cases. Also unexpectedly, removing gender with diarization produces a very similar performance decrease compared with removing a single dimension from the baseline. The most likely reason for this is the nature of the SCOTUS corpus, which is particularly male dominated. Although the verification trials were selected to be speaker balanced (77% male), this is not the case with diarizing the raw test set recordings, which in terms duration are >90% male. Thus when scoring all pairs of segments, the overwhelming majority of pairs cannot benefit from distinguishing by gender.

	EER	Δ%
Baseline	4.22%	-
Baseline (avg. excl. 1 dim)	4.30%	1.8%
Baseline (avg. excl. 10 dim)	4.81%	14.0%
Baseline (avg. excl. 11 dim)	4.88%	15.6%
All dims	6.68%	-
-Gender dim	7.63%	14.2%
-Nationality dims	8.76%	31.1%
-Nationality, Gender dims	10.19%	52.5%

Table 2: *Verification performance on VoxCeleb, using the SplitDim-Adv embeddings and the subset of dimensions. Also shown is the relative percentage increase in EER compared to using all dimensions. -Gender removes 1 dim and -Age removes 10 dims.*

	EER	Δ%	DER	Δ%
Baseline	2.10%	-	32.19%	-
Baseline (excl. 1-d)	2.13%	1.4%	32.76%	1.77%
Baseline (excl. 10-d)	2.40%	14.1%	36.69%	14.0%
Baseline (excl. 11-d)	2.44%	16.2%	37.48%	16.4%
All dims	3.52%	-	29.74%	-
-Gender dim	3.67%	4.26%	30.26%	1.75%
-Age dims	4.41%	25.3%	35.07%	17.9%
-Age, Gender dims	4.62%	35.1%	35.69%	20.0%

Table 3: *Verification and diarization performance on SCOTUS, using the SplitDim-Adv embeddings. -Gender removes 1 dim and -Age removes 10 dims.*

6. Conclusions

In this work, we showed that utilizing multi-task learning alongside adversarial training can effectively disentangle and factorize speaker attributes in the speaker embedding space, with the use of the adversaries essential in separating out sources of variation. Using these disentangled representations, we looked at how gender, age and speaker nationality contribute toward speaker separability, finding that gender information was a significant source of information when discerning between speakers in the embedding space for verification, compared to that of nationality or age.

7. Acknowledgements

Supported by an EPSRC iCASE studentship collaboration with the BBC.

8. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: robust DNN embeddings for speaker recognition," in *IEEE ICASSP*, 2018, pp. 5329–5333. [Online]. Available: http://www.openslr.org.http://danielpovey.com/files/2018_icassp_xvectors.pdf

[2] G. Sell, D. Snyder, A. Mccree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, and S. Khudanpur, "Diarization is Hard: some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge," in *Interspeech*, 2018, pp. 2808–2812.

[3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[4] M. Jessen, "Speaker classification in forensic phonetics and acoustics," *Lecture Notes in Computer Science*, vol. 4343 LNAI, pp. 180–204, 2007. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-74200-5{_}10

[5] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[6] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, "Towards a definition of disentangled representations." [Online]. Available: http://arxiv.org/abs/1812.02230

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives." [Online]. Available: http://arxiv.org/abs/1206.5538

[8] J. J. DiCarlo and D. D. Cox, "Untangling invariant object recognition," vol. 11, no. 8, pp. 333–341. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661307001593

[9] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations." [Online]. Available: http://arxiv.org/abs/1811.12359

[10] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," *Interspeech*, vol. 2017-August, pp. 2616–2620, 2017.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Interspeech*, vol. 2018-Septe, no. ii, pp. 1086–1090, 2018.

[12] J. Williams and S. King, "Disentangling style factors from speaker representations," in *Interspeech*, vol. 2019-Septe, 2019, pp. 3945–3949. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1769

[13] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-Vectors Meet Emotions: A Study On Dependencies Between Emotion and Speaker Recognition." Institute of Electrical and Electronics Engineers (IEEE), apr 2020, pp. 7169–7173.

[14] S. Maiti, E. Marchi, and A. Conkie, "Generating Multilingual Voices Using Speaker Space Translation Based on Bilingual Speaker Data," in *IEEE ICASSP*, vol. 2020-May, 2020, pp. 7624–7628.

[15] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the Information Encoded in X-Vectors," in *IEEE ASRU*, 2019, pp. 726–733. [Online]. Available: https://github.com/Kyubyong/g2p

[16] C. Luu, P. Bell, and S. Renals, "Leveraging speaker attribute information using multi task learning for speaker verification and diarization." [Online]. Available: http://arxiv.org/abs/2010.14269

[17] R. Aloufi, H. Haddadi, and D. Boyle, "Privacy-preserving voice analysis via disentangled representations," pp. 1–14. [Online]. Available: http://arxiv.org/abs/2007.15064

[18] A. Nautsch, J. Patino, N. Tomashenko, J. Yamagishi, P.-G. Noe, J.-F. Bonastre, M. Todisco, and N. Evans, "The privacy ZEBRA: Zero evidence biometric recognition assessment," pp. 1698–1702. [Online]. Available: http://arxiv.org/abs/2005.09413

[19] J. Williams, J. Yamagishi, P.-G. Noe, C. V. Botinhao, and J.-F. Bonastre, "Revisiting speech content privacy." [Online]. Available: http://arxiv.org/abs/2110.06760

[20] 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." [Online]. Available: http://arxiv.org/abs/2012.04454

[21] 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." [Online]. Available: http://arxiv.org/abs/2110.05840

[22] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, U. Dogan, M. Kloft, F. Orabona, and T. Tommasi, "Domain-adversarial training of neural networks," in *Advances in Computer Vision and Pattern Recognition*, 5 2015, vol. 17, no. 9783319583464, pp. 189–209. [Online]. Available: https://arxiv.org/pdf/1505.07818.pdfhttp://arxiv.org/abs/1505.07818

[23] Y. Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *Interspeech*, 2016, pp. 2369–2372.

[24] Transcripts and recordings of oral arguments - supreme court of the united states. [Online]. Available: https://www.supremecourt.gov/oral_arguments/availabilityoforalargumenttranscripts.aspx

[25] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," Tech. Rep., 2008. [Online]. Available: https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf