The Speaker and Language Recognition Workshop (Odyssey 2020)
1-5 November 2020, Tokyo, Japan

An empirical analysis of information encoded in disentangled neural speaker representations

Raghuveer Peri, Haoqi Li, Krishna Somandepalli, Arindam Jati, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory,
University of Southern California,
Los Angeles, CA, USA.
{rperi, haoqili, somandep, jati}@usc.edu, shri@ee.usc.edu

Abstract

The primary characteristic of robust speaker representations is that they are invariant to factors of variability not related to speaker identity. Disentanglement of speaker representations is one of the techniques used to improve robustness of speaker representations to both intrinsic factors that are acquired during speech production (e.g., emotion, lexical content) and extrinsic factors that are acquired during signal capture (e.g., channel, noise). Disentanglement in neural speaker representations can be achieved either in a supervised fashion with annotations of the nuisance factors (factors not related to speaker identity) or in an unsupervised fashion without labels of the factors to be removed. In either case, it is important to understand the extent to which the various factors of variability are entangled in the representations. In this work, we examine speaker representations with and without unsupervised disentanglement for the amount of information they capture related to a suite of factors. Using classification experiments we provide empirical evidence that disentanglement reduces the information with respect to nuisance factors from speaker representations, while retaining speaker information. This is further validated by speaker verification experiments on the VOiCES corpus in several challenging acoustic conditions. We also show improved robustness in speaker verification tasks using data augmentation during training of disentangled speaker embeddings. Finally, based on our findings, we provide insights into the factors that can be effectively separated using the unsupervised disentanglement technique and discuss potential future directions.

1. Introduction

Speaker embeddings are low-dimensional representations of speech that capture speaker characteristics. They have numerous applications in tasks involving automatic speaker recognition, such as speaker diarization (identifying who spoke when) [1], voice biometrics [2], anti-spoofing [3] and personalized services such as in smart home devices [4]. Real-world speaker recognition requires that speaker embeddings capture speaker characteristics robust to all other attributes unrelated to the speaker's identity, such as the acoustic conditions, microphone characteristics and (aspects of) lexical content in the speech signal.

The topic of extracting robust speaker representations invariant to various factors of variability has been widely studied in the literature. A powerful joint factor analysis (JFA) based approach was developed in [5], where speaker 'supervectors' were factored into speaker-independent, speaker-dependent, channel-dependent and residual factors. The goal was to separately model the different factors of variability. But it was found that the channel-dependent factors, which were expected to capture only the characteristics of the transmission channel, also contained speaker-related information [6]. To overcome this challenge, a total variability modeling (TVM) approach was proposed [7], where no distinction was made between the speaker and session variability factors. These speaker embeddings, called i-vectors are further processed through additional channel compensation steps in the total variability space to minimize session variability [8, 9]. They have been shown to perform well on speaker verification tasks.

Recently, supervised speaker modeling techniques have been developed [10, 11]. These methods differ from the previous approaches in that they do not try to explicitly separate the factors of variability. The speaker representations proposed in [11], called x-vectors, are extracted from the bottleneck layer of a time-delay neural network, which was trained on a large corpus of augmented audio recordings to recognize speaker identity. They have been shown to outperform i-vectors for speaker verification, especially in utterances that are shorter than 10 seconds [12]. Systems employing x-vectors have achieved state-of-the-art performance in applications such as speaker verification [11] and speaker diarization [13]. Since x-vector systems were not trained to explicitly remove factors of variability, they retain information unrelated to the speaker identity [14, 15].

More recently, other approaches to obtain robust speaker embeddings have been proposed, where models were trained to induce robustness to specific factors of variability in a supervised fashion using additional labels of channel conditions [16, 17]. However, it is not practical to obtain such labelled data in real-world scenarios where the recording conditions or communication context are often unknown. In order to overcome this challenge, we recently proposed a method to disentangle speaker-related factors from the other factors unrelated to speaker identity without prior knowledge of the channel conditions [18]. The speaker embeddings proposed in [18] were obtained from x-vectors using an adversarial invariance induction technique. We showed improvements on speaker verification tasks using this approach of disentangling speaker-related factors from nuisance factors in the embeddings, without explicit supervision of the factors to be disentangled.

However, analysis of the effect of disentanglement on speaker representations has been under-explored. It is important to understand the extent to which the various factors are entangled in these representations to compare the performance of different speech embeddings for downstream tasks. Such analyses also provide valuable directions to improve the task-specific robustness of speech representations to various confounding fac-

10.21437/Odyssey.2020-28

Table 1: Factors in speaker embeddings and corpora used

	Factors considered	Study corpora
Channel factors	Mic Noise Room	VOiCES [19]
Content factors	Emotion/setiment Language Lexical	IEMOCAP [20], MOSEI [21] Mozilla [22] RedDots [23]
Speaker factors	Speaker identity Gender	IEMOCAP, VOiCES, RedDots IEMOCAP

tors. In this work, we aim to understand the effect of disentanglement on the amount of information retained in the speaker embeddings with respect to various factors.

The contributions of this work are as follows:

- Identify speech datasets to analyze the information encoded in speaker representations with respect to a suite of possible factors including speaker identity, gender, lexical content, emotion, sentiment, noise-type and recording channel.
- Assess the extent to which the speaker-related factors and factors unrelated to speaker identity are encoded in speaker embeddings with and without disentanglement, where disentanglement is achieved using the method proposed in [18].
- Show benefits of augmenting the data used for training the disentanglement models in further improving the robustness of speaker recognition.

2. BACKGROUND

Several works have studied the effect of intrinsic factors which are acquired by the signal during the speech production stage, such as the emotional content, lexical information, language [24–26]. Other studies have focused on extrinsic factors that are acquired at the signal recording stage [27]. These different factors of variability are entangled to varying degrees in the speaker representations.

Following [28], for ease of analysis, we categorize the factors into three distinct categories: *channel factors* which are encoded in the speech signal during the process of signal acquisition, such as background acoustic noise, microphone characteristics, room response and acoustic scene, *content factors* which encode information about the spoken content including its prosodic content, such as the lexical aspects, emotion, sentiment, language and accent [29], *speaker factors* which are inherent to the speaker, such as speaker identity, age and gender.

In Table 1 we summarize few of the exemplar factors belonging to these three categories. We also list the corpora that we have considered in this paper. In the following sections, we discuss the 3 different factors of variability in detail.

Channel factors: These factors have been extensively studied in literature in the context of speaker recognition. The goal of robust speaker recognition systems is to rid the speaker embeddings of these factors, while retaining the speaker-related factors. As mentioned in Section 1, early speaker recognition systems either attempted to decompose channel-dependent factors from the speaker-dependent factors during speaker modeling [5], or introduced channel compensation steps in the total variability space [9]. X-vector based methods do not make this separation explicit and can contain significant amount of channel-related information, as shown through channel classification experiments [14].

Content factors: They encode information about the spoken content including the prosodic variations in speech. Examples include lexical content, emotional content, sentiment, language identity etc. These factors have been shown in past works to be entangled with speaker-related factors in speaker representations. For example, examining the amount of phoneme-level information present in speaker embeddings, it was found that better performance in phoneme classification tasks does not translate to improved speaker recognition performance [30]. Furthermore, speaker embeddings that perform well for speaker recognition tasks have been shown to capture segment-level characteristics rather than low-level phoneme information [31].

It was found in past literature that both i-vectors and x-vectors capture speaker-related factors along with significant information related to the other factors [14,28]. However, it is important to note that the inter-dependence between factors arising due to the construction of the dataset were not considered in [14]. For example, lexical content information was investigated using sentence classification task, but the data included sentences that were unique to particular speakers, making it challenging to disambiguate the speaker information from the lexical information using only the sentence classification performance. Williams et al. [15], examined x-vectors for the task of subject-independent emotion recognition. Their experiments showed that x-vectors trained for speaker recognition task could predict the emotion content in speech even without any additional supervision using emotion labels.

Speaker factors: These factors are inherent to the speaker (c.f., demographics), and capture the identity of the speaker to various degrees. For example, while gender and age are not sufficient to fully recognize a speaker's identity from speech, robust speaker embeddings can effectively capture these dimensions. Thus examining how speaker embeddings perform for identifying these individual factors is key to understanding robustness for speaker recognition. In particular, it has been found that i-vectors and x-vectors can be successfully used to classify gender [14,28]. However, it is still unclear whether speaker embeddings that are more discriminative of gender significantly improve speaker recognition performance. In this paper we also assess these factors when the labels are available.

We probe into the information encoded in two speaker embeddings based on the work in [18] and compare with that of x-vectors. We show through speaker verification experiments on the VOiCES dataset [19] that the speaker embeddings after disentanglement retain information pertaining to speaker identities. Since the speaker representations extracted using the technique in [18] do not rely on labelled information about the nuisance factors, we hypothesize that such disentanglement results in speaker representations that contain minimal information related to all the nuisance factors, while retaining speaker-related information. We chose classification tasks related to a particular factor of variability as a proxy to the amount of information encoded in the speaker representations related to that factor as has been done in several previous works [15,28,31,32]. Results suggest that the process of disentanglement reduces the amount of information present in speaker embeddings pertaining to factors not related to the speaker identity.

3. METHODOLOGY

3.1. Baseline

We extract x-vectors using the publicly available pre-trained model [1] and consider them as the baseline speaker embeddings trained without disentanglement. The x-vector model was trained on a large corpus of audio recordings to predict speakers. These audio recordings were further augmented by artificially adding background noise and music at varying signal-to-noise ratio levels [11]. In order to simulate the effect of reverberation, the audio signals were convolved with various room impulse responses. As discussed in Section 2, since x-vectors were not trained with an explicit disentanglement stage, they retain information pertaining to factors unrelated to speaker identity.

3.2. Disentangled speaker embeddings

We employ the unsupervised adversarial invariance (UAI) technique, which was originally proposed in [33] and modified to disentangle the speaker factors from the other factors present in x-vectors [18]. The central idea behind the UAI technique is to project the input x-vectors into a split representation consisting of two embeddings, referred to as $\mathbf{h}_1$ and $\mathbf{h}_2$. While $\mathbf{h}_1$ is trained with the objective of capturing speaker-specific information, $\mathbf{h}_2$ is trained to capture all other nuisance factors. This is done by training two models in an adversarial fashion. The goal of one model, called main model, is to predict speakers using $\mathbf{h}_1$ as input and reconstruct the x-vectors using $\mathbf{h}_2$ as input. The second model, called the adversarial model, is trained to minimize the predictive power of $\mathbf{h}_1$ from $\mathbf{h}_2$ and vice-versa. The two models are adversarially trained in a minimax game fashion. The speaker prediction task forces $\mathbf{h}_1$ to capture speaker-related information, while the reconstruction task ensures that $\mathbf{h}_2$ captures information related to all factors. Further, a noisy version of $\mathbf{h}_1$ is used during the reconstruction task along with $\mathbf{h}_2$, so that the network learns to treat $\mathbf{h}_1$ as an unreliable source of information for the reconstruction, hence ensuring $\mathbf{h}_1$ does not contain information about factors other than the speaker. Detailed explanation of the UAI method to disentangle speaker representations can be found in [18].

We trained two models using the UAI technique that differ in the training data used:

- M1: Trained without artificially augmented data
- M2: Trained with artificially augmented data

For consistency, we use the same model as proposed in [18], denoted M1. We developed the model M2 to investigate the benefit of artificially augmenting the training samples on the robustness of speaker embeddings.

3.3. Factor prediction

As mentioned in Section 2, to measure the extent of entanglement of the various factors in speaker embeddings, we designed classification experiments. In these experiments, the speaker embeddings were used as input to predict each factor of variability. Such a method of analysis assumes that if a factor is encoded in the speaker embeddings, a classifier can be trained to predict the factor using the speaker representations as input. Further, the classification accuracy can be considered a proxy for how well the factor has been encoded in these speaker embeddings [28].

[1] https://kaldi-asr.org/models/m7

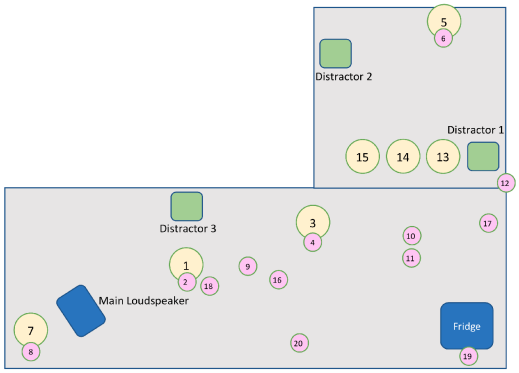

Figure 1: Example room configuration in VOiCES dataset [34]. Distractor represents noise source and circles represent microphones

4. DATASETS

As mentioned in Section 2, we performed experiments on a number of publicly available datasets. Each dataset was chosen to enable the exploration of individual factors, while providing an analysis of how these factors are manifested in speaker representations on real-world data.

VOiCES: Recordings collected from 4 different rooms with microphones placed at various fixed locations, while a loudspeaker played clean speech samples from the Librispeech [35] dataset. Along with speech, noise was played back from loudspeakers present in the room, to simulate real-life recording conditions. Figure 1 shows one such room configuration and data collection setup where "Distractor" represents noise source and the circles represent the available microphones. This dataset was chosen for the availability of annotations regarding the acoustic conditions including noise types and microphone locations. We use the VOiCES corpus to evaluate speaker verification performance under various acoustic conditions. We perform our evaluations on the phase-2 release of the dataset to be consistent with the data used during the evaluation in VOiCES challenge [34]. This subset consists of recordings from 2 rooms collected using 20 microphones and contains 4 distinct noise types. We also use this dataset to explore the amount of information related to the speaker, room, noise and microphone type.

IEMOCAP: A multi-modal corpus consisting of video, audio, face motion, hand movements and text transcriptions. These modalities were recorded from 10 actors during scripted and enacted conversations. Annotations are provided for discrete and continuous emotion ratings. Following [15], we make use of a small subset consisting of the audio portion of the data with a single emotion label per utterance, leading to 1403 utterances corresponding to 4 discrete emotions, angry, sad, happy and neutral. This dataset provides the capability to investigate the amount of emotional information present in the speaker embeddings. Further, we use this dataset to perform speaker identification experiments to explore the speaker information present in the embeddings.

MOSEI: An audio dataset of more than 23000 YouTube videos annotated for emotion and sentiment. Audio from this corpus was used for the analysis of emotion and sentiment information captured in speaker embeddings. Since, the emotions elicited

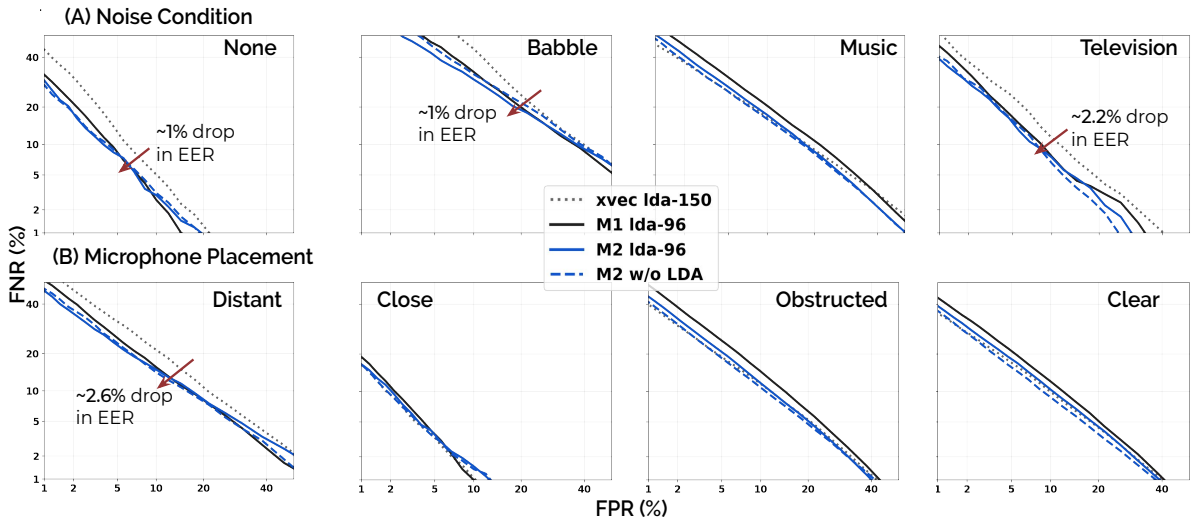

Figure 2: DET curves of speaker verification task using different speaker embeddings with and without disentanglement in several (A) Noise conditions and (B) Microphone placements.

in all the recordings were spontaneous, this database was useful to validate the generalizability of our analysis to real-world spontaneous data. We performed pre-processing of the raw labels provided in the corpus. Each audio recording in the dataset was annotated for scores corresponding to six emotion labels. We chose the emotion label with the maximum score for each audio recording. For sentiment analysis, we convert the labels provided in the dataset into 3 distinct labels, positive, negative and neutral sentiment.

Mozilla: A publicly available corpus of crowdsourced audio recordings. It consists of transcribed text from multiple languages, from which we chose a subset of 4 languages (English, German, Mandarin and Turkish) for ease of analysis. The languages are chosen such that sufficient number of samples exist for each language. This corpus is used for the language prediction task.

RedDots: Speech recordings collected from participants using their own audio recording device, typically a mobile phone. The dataset comprises recordings of participants reading out sentences, some of which were common across all the participants while the others were unique to each participant. To ensure that experiments involving lexical content factor were not confounded by variability due to speakers, we prune the data and obtain a small subset of recordings consisting of only the 10 common sentences spoken by all the speakers. This corpus is used for analysis of lexical content and also for the gender prediction task.

Vox_clean: To ensure fair comparison with x-vectors baseline, we trained the models in [18] using a subset of the data that was used to develop the pre-trained x-vector models. It consists of a combination of the development and test splits of VoxCeleb2 [36] and the development split of VoxCeleb1 [37] datasets. Since the dataset is sourced from unconstrained recording conditions, there exists a huge amount of variability in terms of channel conditions and the content factors such as lexical and emotion content [38]. It consists of audio recordings corresponding to $1.2M$ utterances from 7323 distinct speakers. We refer to this subset as Vox_clean. We also augment this data by artificially adding noise following [11], using music and noise samples from the MUSAN corpus [39]. Artificial reverberation was simulated using room impulse response samples from `https://www.openslr.org/`. We refer to this augmented dataset as **Vox_aug**. This process doubled the amount of training data to $2.4M$ utterances keeping the number of speakers the same.

5. EXPERIMENTS

5.1. Speaker verification

5.1.1. Setup

We performed several experiments on the VOiCES dataset to verify the efficacy of disentanglement in making speaker embeddings robust to various factors of variability. Following [19] and [27], we experimented with several combinations of noise types and microphone locations to understand the effect of each factor on speaker verification performance. In a similar fashion as [18], we consider two distinct factors, noise conditions: none, babble, television and music, and microphone location: distant, close, clear and obstructed. By abuse of notation, we denote the disentangled embeddings by the same name as the model from which they were extracted, i.e., M1 and M2.

We apply dimensionality reduction using linear discriminant analysis (LDA) and score the speaker verification trials using a probabilistic linear discriminant analysis (PLDA) backend for all the embeddings obtained with and without disentanglement. The LDA and PLDA models were learnt on the training data for our proposed system, while for the baseline x-vector system we used the pre-trained models. For the embeddings extracted after disentanglement (both M1 and M2), we use a dimension of 96 after LDA (chosen based on speaker verification performance on the VOiCES challenge development portion), while for x-vectors we use 150 as the reduced dimension to be compatible with the pre-trained PLDA model. We also experimented with removing the LDA-based dimensionality reduction stage on the disentangled speaker embeddings.

Table 2: Accuracy (%) of speaker embeddings to classify different factors. Robust speaker embeddings are expected to perform poorly for classifying non-speaker related tasks. The best model for each factor with respect to this criterion is shown in bold

	Factors	Dataset	x-vector (512 dim)	x-vector + PCA (128 dim)	Ours: M1 (128 dim)
Channel factors	Room	VOiCES	99.8	99.7	**97.3**
	Mic		91.0	83.3	**57.4**
	Noise		94.9	92.2	**76.2**
Content factors	Emotion	IEMOCAP	91.5	91.7	**80.9**
		MOSEI	62.4	61.5	**60.4**
	Sentiment	MOSEI	53.9	53.0	**49.3**
	Lexical	RedDots	98.0	97.1	**80.3**
	Language	Mozilla	97.2	96.6	**95.3**
Speaker factors	Speaker	VOiCES	**99.6**	99.6	98.5
		IEMOCAP	80.1	**81.8**	77.8
	Gender	IEMOCAP	**98.9**	98.0	97.2
		RedDots	99.0	**99.5**	97.6

5.1.2. Results

Figure 2 shows the detection error trade-off (DET) plots of the various speaker verification experiments with False Positive Rate (FPR) on the x-axis and False Negative Rate (FNR) on the y-axis. Each subplot compares four speaker embeddings, x-vector reduced to 150 dimensions using LDA (**xvec lda-150**), M1 reduced to 96 dimensions (**M1 lda-96**), M2 reduced to 96 dimensions (**M2 lda-96**) and M2 without LDA-based dimensionality reduction (**M2 w/o lda**).

The first row of subplots in Figure 2 shows the DET curves across different noise conditions. We observe that in majority of the cases disentangled speaker embeddings obtain lower error rates at most operating points. In particular, for the television and babble noise conditions, which are considered challenging due their speech-like characteristics [27], we observed a respective absolute reduction of 2.2% and 1% in the equal error rate (EER). Furthermore, we observe that for the distant microphone scenario, speaker embeddings after disentanglement provide lower error rates at almost all operating points, with a 2.6% reduction in EER.

5.2. Probing analysis

5.2.1. Setup

As mentioned in Section 2, we analyze the extent of information present in the speaker embeddings with respect to various factors. For each factor that was analyzed, we extracted both x-vectors and disentangled speaker embeddings (denoted M1) from the corresponding dataset for that factor. We also reduce the dimension of x-vectors (denoted xvector+PCA) using principal component analysis (PCA), similar to [15]. This was done to match the dimension of x-vectors with that of the disentangled speaker embeddings, to ensure fair comparison with respect to embedding dimension.

We trained a classification model for each factor using a simple feed-forward deep neural network. The classifier contains of 4 fully connected layers with 256 hidden neurons and ReLU activation function in between. Similar to [15], we use L2 regularization, Adam optimizer with learning rate lr = 0.0002, and an early-stopping criterion monitored by the loss on validation set. The classifier models were trained to predict the factor with the speaker embeddings as input.

For the IEMOCAP dataset, for fair comparison, we use the same train and test split as in [15]. For the RedDots dataset, in the lexical content analysis task, for each speaker we randomly split 80% of the data into training, 10% into validation and 10% as test part. In the gender classification task, based on the previous split, we further perform minority class sub-sampling to balance the data for each gender.

Results summarizing the classification performance of the speaker embeddings with and without disentanglement are reported in Table 2, where the accuracy (%) values were rounded to the nearest decimal.

5.2.2. Channel factors

We explore three different channel variability conditions: microphone id, noise and room type. We trained classifiers to predict the microphone at which the recording was collected given the speaker embeddings extracted from a recording. We built classifiers to predict the noise type from the speaker embeddings. We also trained classification models to predict the room in which the recordings were made, using speaker embeddings as input.

Results

We observe from Table 2 that there is a fairly large difference in the predictive power of x-vectors and disentangled speaker embeddings (M1) when classifying microphones and noise types. This suggests that the disentanglement method is able to successfully reduce the channel information from speaker representations. However, all embeddings perform well in predicting the room. It is unclear if this behavior is due to an unknown factor of variability between the rooms.

5.2.3. Content factors

Experiments were performed to analyze the information related to four different content factors, lexical content, emotional content, sentiment and language identity using the corresponding corpora mentioned in Table 1.

Results

Results on lexical content prediction show an 18% reduction in classification accuracy upon disentanglement of x-vectors, suggesting the successful reduction in lexical information from

speaker embeddings. Results for emotion classification show a 10% reduction in accuracy on the IEMOCAP dataset. The results on MOSEI dataset show a similar trend of reduced accuracy in M1 for the emotion classification tasks, though the difference is not as substantial. For the sentiment classification task, there is an absolute 5% reduction in accuracy after disentanglement. Finally, with respect to language prediction, all the embeddings perform well, with a minor decrease in accuracy when disentangled representations are used.

5.2.4. Speaker factors

To analyze the speaker-related information present in the speaker representations, we performed separate experiments to predict the speaker identity and speaker gender. We used the speaker labels in the VOiCES dataset from which 10 speakers were randomly chosen for the speaker identification task. We trained classifiers to predict the speaker identity given the speaker embeddings. We also performed similar classification experiments on the IEMOCAP dataset. We also trained classifiers on the RedDots dataset to make binary gender predictions using the speaker embeddings as input .

Results

For the speaker id task on the VOiCES dataset, we observe that M1 performs almost on par with x-vectors, and close to 100% accuracy, suggesting that speaker identity is retained during the disentanglement process. This result complements the speaker verification performance, see Section 5.1.

We further observe a similar trend for the gender prediction task on the RedDots dataset, where all the speaker embeddings achieve close to 100% accuracy. This is consistent with past work [14, 28].

However, results on IEMOCAP did not follow the same trend, though the differences in performance are small. The reason for this could be the implicit co-dependence between the emotions and the speaker labels in the dataset. In order to test this, we conducted hypothesis testing on the contingency table using a chi-squared test[2] for the sample frequencies of the emotion and speaker variables in the IEMOCAP dataset. This test ($p \ll 0.001$) revealed that the two variables were dependent in this dataset. This makes it difficult to disambiguate the effects of one factor on the other. Furthermore, we noticed that a few audio recordings contained overlapping speech leading to noisy speaker labels.

5.3. Discussion

We offer the following insights based on our evaluations using x-vectors and disentangled representations:

- Channel factors can be effectively removed from speaker embeddings using unsupervised disentanglement. It further improves robustness of the embeddings for speaker recognition tasks in challenging acoustic conditions. This is consistent with our previous findings
- The emotional content information is also minimized from speaker embeddings during disentanglement, even though emotion labels were not used during training. This can be explained based on the findings in [38], where it was shown that audio recordings in the Voxceleb dataset contain expressive speech.
- PCA on x-vectors to reduce dimensionality does not have a significant effect on the prediction performance. This suggests that explicit disentanglement is beneficial over simple dimensionality reduction techniques.
- Disentanglement of speaker embeddings retains gender-related information. This suggests that speaker gender, as captured in these representations, is a crucial part of the speaker's identity. If removal of gender factor from speaker representations is desired, then other joint modeling approaches need to be explored. Further, other individual attributes such as age should be investigated in a similar manner.
- Consistent with past literature in this domain, we observe that additional data augmentation further improved the performance on the speaker verification task. However, in the context of disentanglement, data augmentation did not reduce the information of the nuisance factors. In our future work, we wish to investigate if informed data augmentation for individual nuisance factor can help disentanglement.

[2] H_0: the two variables of the contingency table are independent. Reject H_0 if $p < \alpha$ with $\alpha = 0.01$

6. Conclusion

In this work, we showed that disentanglement of neural speaker representations helps improve robustness of speaker embeddings on speaker verification tasks in challenging acoustic conditions. We further showed through experimental evaluation on several datasets and a suite of factors that disentanglement reduces the amount of information encoded in the speaker embeddings that is not related to speaker identity, while retaining speaker-related information. We hope that these findings will offer novel research directions to develop robust speaker recognition systems.

7. References

[1] Homayoon Beigi, "Speaker recognition," in *Fundamentals of Speaker Recognition*, pp. 543–559. Springer, 2011.

[2] Judith A. Markowitz, "Voice biometrics," *Commun. ACM*, vol. 43, no. 9, pp. 66–73, Sept. 2000.

[3] Chunlei Zhang, Chengzhu yu, and John Hansen, "An investigation of deep learning frameworks for speaker verification anti-spoofing," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, pp. 1–1, 01 2017.

[4] Dong-Gyu Shin and Moon-Seog Jun, "Home iot device certification through speaker recognition," 07 2015, pp. 600–603.

[5] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[6] Najim Dehak, *Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification*, Ph.D. thesis, 2009, AAINR50490.

[7] 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, May 2011.

[8] Alex Solomonoff, William M Campbell, and Carl Quillen, "Nuisance attribute projection," *Speech Communication*, pp. 1–73, 2007.

[9] Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1969–1978, 2007.

[10] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[11] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[12] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Interspeech*, 2017, pp. 999–1003.

[13] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4930–4934.

[14] Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur, "Probing the information encoded in x-vectors," *arXiv preprint arXiv:1909.06351*, 2019.

[15] Jm Williams and Simon King, "Disentangling style factors from speaker representations," in *INTERSPEECH 2019*, 2019.

[16] G. Bhattacharya, J. Monteiro, J. Alam, and P. Kenny, "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6226–6230.

[17] Jianfeng Zhou, Tao Jiang, Lin Li, Qingyang Hong, Zhe Wang, and Bingyin Xia, "Training multi-task adversarial network for extracting noise-robust speaker embedding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6196–6200.

[18] Raghuveer Peri, Monisankha Pal, Arindam Jati, Krishna Somandepalli, and Shrikanth S. Narayanan, "Robust speaker recognition using unsupervised adversarial invariance," *ArXiv*, vol. abs/1911.00940, 2019.

[19] Colleen Richey, Maria A Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, et al., "Voices obscured in complex environmental settings (voices) corpus," *arXiv preprint arXiv:1804.05053*, 2018.

[20] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[21] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018, pp. 2236–2246, Association for Computational Linguistics.

[22] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[23] Kong Aik Lee, Anthony Larcher, Guangsen Wang, Patrick Kenny, Niko Brümmer, David van Leeuwen, Hagai Aronowitz, Marcel Kockmann, Carlos Vaquero, Bin Ma, et al., "The reddots data collection for speaker recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] Juliette Kahn, Nicolas Audibert, Solange Rossato, and Jean-François Bonastre, "Intra-speaker variability effects of speaker verification performance," *Odyssey-2010, the speaker and Language recognition Workshop*, 06 2010.

[25] S. Parthasarathy, C. Zhang, J. H. L. Hansen, and C. Busso, "A study of speaker verification performance with expressive speech," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5540–5544.

[26] Huanjun Bao, Ming-Xing Xu, and Thomas Fang Zheng, "Emotion attribute projection for speaker recognition on emotional speech," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[27] Mahesh Kumar Nandwana, Julien van Hout, Mitchell McLaren, Allen R Stauffer, Colleen Richey, Aaron Lawson, and Martin Graciarena, "Robust speaker recognition from distant speech under real reverberant environments using speaker embeddings.," in *Interspeech*, 2018, pp. 1106–1110.

[28] Shuai Wang, Yanmin Qian, and Kai Yu, "What does the speaker embedding encode?," in *Interspeech*, 2017, pp. 1497–1501.

[29] 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)*, April 2018, pp. 5094–5098.

[30] Alicia Lozano-Diez, Anna Silnova, Pavel Matejka, Ondrej Glembek, Oldrich Plchot, Jan Pesan, Lukas Burget, and Joaquin Gonzalez-Rodriguez, "Analysis and optimization of bottleneck features for speaker recognition.," in *Odyssey*, 2016, pp. 352–357.

[31] Suwon Shon, Hao Tang, and James R. Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1007–1013, 2018.

[32] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg, "Fine-grained analysis of sentence embeddings using auxiliary prediction tasks," 2016.

[33] Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan, "Unsupervised adversarial invariance," in *Advances in Neural Information Processing Systems*, 2018, pp. 5092–5102.

[34] Mahesh Kumar Nandwana, Julien Van Hout, Mitchell McLaren, Colleen Richey, Aaron Lawson, and Maria Alejandra Barrios, "The voices from a distance challenge 2019 evaluation plan," *arXiv preprint arXiv:1902.10828*, 2019.

[35] 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)*, April 2015, pp. 5206–5210.

[36] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[37] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[38] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 292–301.

[39] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *ArXiv*, vol. abs/1510.08484, 2015.