Speech Disentanglement for Analysis and Modification of Acoustic and Perceptual Speaker Characteristics

Frederik Rautenberg[1], Michael Kuhlmann[1], Janek Ebbers[1], Jana Wiechmann[2], Fritz Seebauer[2], Petra Wagner[2], Reinhold Haeb-Umbach[1]

[1] *Department of Communications Engineering, Paderborn University Email: {rautenberg, kuhlmann, ebbers, haeb}@nt.upb.de*

[2] *Phonetics Work Group, Bielefeld University Email: {jana.wiechmann, fritz.seebauer, petra.wagner}@uni-bielefeld.de*

Abstract

Popular speech disentanglement systems decompose a speech signal into a content and a speaker embedding, where a decoder reconstructs the input signal from these embeddings. Often, it is unknown, which information is encoded in the speaker embeddings. In this work, such a system is investigated on German speech data. We show that directions in the speaker embeddings space correlate with different acoustic signal properties that are known to be characteristics of a speaker, and manipulating these embeddings in that direction, the decoder synthesises a speech signal with modified acoustic properties.

Introduction

With the advent of deep learning, speech processing has reached an unprecedented level of performance. Just like any other statistical model class, deep neural networks can be categorized into discriminative and generative models. While classification tasks are traditionally approached by discriminative models, generative models allow for the generation of speech samples. A prominent example is the variational autoencoder (VAE) that has been shown to generate realistic speech samples, in particular, if extended for sequence modeling as is done in the dynamical variational autoencoder. The VAE consists of an encoder and a decoder neural network. The encoder maps the observed data to a latent space that corresponds to the parameters of a normal distribution, which serves as variational approximation to an analytically intractable posterior. Then samples are drawn from the normal distribution which the decoder takes as input to generate new data points.

In [1] we have shown, that by using two encoders instead of one, a speech signal can be disentangled in a completely unsupervised fashion into embedding vectors that capture short-term (fast) variations of the signal and an embedding vector that captures the temporally stable properties of the signal. We call the former content and the latter speaker or style embedding in the following. The rationale behind this is the intuition that fast variations are caused by the linguistic content, while properties of the speaker, the environment or other long-term properties, e.g., emotion, are temporally stable or change only very slowly. With this in mind appropriate loss functions can be designed for the training of this so-called Factorized Variational Autoencoder (FVAE). Further, by exchanging the speaker embeddings with those of another speaker, while leaving the content embeddings untouched, voice conversion can be carried out, as was illustrated in [2].

However, what is exactly encoded in the the speaker embedding, and can the components of the embedding vector be interpreted acoustically or even perceptually? We carry out a statistical analysis of the embedding vectors to find out whether they indeed encode acoustic signal properties that are known to be characteristic of a speaker. To this end, we investigate which acoustic features are encoded in which components of the embedding vector. We take the widely used openSMILE [3] feature set and we choose two features for a case study, the mean of the fundamental frequency and the Hammarberg index. The former is known to encode, among others, the gender of the speaker, while the second, the ratio of the maximum power below 2000 Hz to the maximum power in the range 2000 Hz − 5000 Hz, can be used for emotion recognition, as was shown in [4].

With a Canonical Correlation Analysis (CCA), we determine the direction in the latent speaker space, that has highest correlation with those two acoustic features. We then modify the speaker embedding along those directions and qualitatively evaluate the effect on the synthesized speech signal[1]. Clearly, the analysis shows that acoustic properties are not aligned with the coordinate axes of the latent feature space. While utterances of the same speaker form well-defined clusters in speaker space, indicating that indeed speaker and content-induced variations are well separated, specific acoustic features are not encoded in single dimensions of the speaker space, calling for better disentanglement, as will be investigated in future work.

Factorized Variational Autoencoder

Here, we briefly describe the FVAE that is used for mapping the speech signal into two latent spaces, one capturing the slow (i.e., speaker/style) and one the fast variations (i.e., content) in the speech signal. For a detailed description, the reader is referred to [1]. From a given utterance, log-mel features $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ are extracted, which serve as input signal to the FVAE. To achieve the disentanglement, the model uses two encoders, a speaker encoder and a content encoder, see Figure 1. The speaker encoder extracts an embedding $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_T]$. It is assumed, that the speaker properties are not changing over time, so a Global Average Pooling (GAP) is applied over the time dimension to create the speaker vector $\bar{\mathbf{s}}$.

To avoid that the content encoder embeds speaker in-

[1] Audio examples: `go.upb.de/daga23`

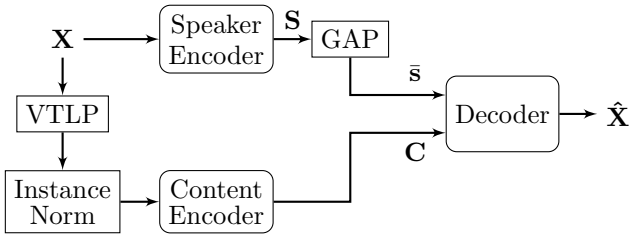

Figure 1: Architecture of the FVAE

formation, the speaker properties in its input are distorted with Vocal Tract Length Perturbation (VTLP) and Instance Normalization. From this preprocessed signal, the content encoder extracts the content embedding $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_T]$. Both embeddings are used by the decoder to reconstruct the input signal $\hat{\mathbf{X}}$, such that the framewise Mean Squared Error (MSE) is minimized

$$L_{\text{rec}} = \frac{1}{T} \sum_{T} ||\hat{\mathbf{x}}_t - \mathbf{x}_t||_2^2 \,. \tag{1}$$

To push the disentanglement further, Contrastive Prediction Coding (CPC) is applied on speaker and content. We assume that the speaker does not change over time. For this reason, the speaker embedding $\mathbf{s}_{t+\tau}$ must be similar to $\mathbf{s}_t$, where τ is the number of steps into the future. But these embeddings must be different to embeddings from other utterances. This is described by the contrastive prediction loss

$$L_{\text{cpc}}^{(S)} = -\frac{1}{T-\tau} \sum_{t=\tau+1}^{T} \frac{\exp(\mathbf{s}_t^{\mathrm{T}} \cdot \mathbf{s}_{t-\tau})}{\sum_{\mathcal{B}} \exp(\tilde{\mathbf{s}}_t^{\mathrm{T}} \cdot \mathbf{s}_{t-\tau})} \,, \tag{2}$$

where $\mathcal{B}$ is a minibatch of speaker embeddings from other speakers. A similar approach is applied to the content embeddings. Here, $\mathbf{c}_{t+\tau}$ and $\mathbf{c}_t$ should not contain any mutual information and to achieve this, the CPC loss is maximized for the content. This gives the overall loss of

$$L = L_{\text{rec}} + \lambda_s L_{\text{cpc}}^{(S)} - \lambda_c L_{\text{cpc}}^{(C)} \,. \tag{3}$$

Acoustic features capturing speaker characteristics

Our goal is to manipulate the voice characteristics of a speaker. Some of these characteristics can be captured by acoustic features. Consequently, if some of the acoustic features of a speech signal are manipulated, then the properties of the speaker are also changed. OpenSMILE [3] is a toolkit to extract acoustic features y for the analysis and classification of speech signals. In this work, we are focusing on two acoustic features, the utterance-wise mean of the pitch y_{p}, which correlates strongly with gender, and the utterance-wise mean of the Hammarberg index y_{h}, which can be used for emotion recognition [4].

Canonical Correlation Analysis

Figure 2 shows the speaker embeddings $\bar{\mathbf{s}}$ of 15 speakers with 4 different utterances each, which are extracted from a pretrained FVAE. The colours represent the speaker IDs or the gender labels. The plot shows, that speaker embeddings from different utterances, but same speaker,

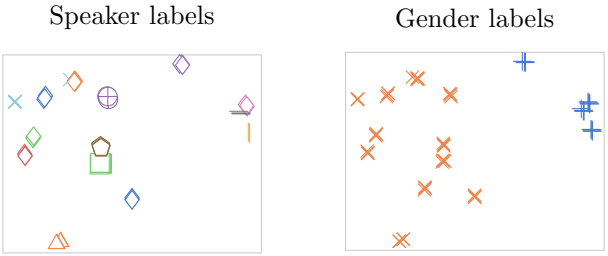

Figure 2: Speaker embeddings $\bar{\mathbf{s}}$ of 15 speakers (unseen in training of FVAE) with 4 utterances each, visualized with t-SNE. In the left picture the embeddings are colored by the speaker ID and on the right by gender.

form clusters, and that the embeddings are clustered according to gender.

It is unclear, though, whether these embeddings also form clusters regarding a speaker-related acoustic feature. To determine the correlation between the embedding and the acoustic feature, we carry out a CCA. CCA is used to find a linear projection $u = \hat{\mathbf{a}}^{\mathrm{T}} \bar{\mathbf{s}}$, which has the maximum correlation with an investigated acoustic feature y, so it follows

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmax}}\, \rho\left(\mathbf{a}^{\mathrm{T}} \bar{\mathbf{s}}, y\right) \,, \tag{4}$$

where u is the canonical variable and ρ is the Pearson correlation coefficient.

Speaker embedding manipulation

The idea of the FVAE is to disentangle a speech signal in a content and a speaker embedding, and the decoder reconstructs the input speech signal from these two embeddings. If the speaker embedding is now replaced by that of another speaker, a speech signal with the same content, but different speech characteristics is generated. Which speech characteristics this speech signal has, depends on the location of the embedding in the latent speaker space. With (4), the projection vector $\hat{\mathbf{a}}$, i.e., the linear combination of components of $\bar{\mathbf{s}}$, is found in the latent speaker space, that correlates most with the acoustic feature y. Manipulating the speaker embedding in the direction of the maximum correlation and using this embedding as input for the decoder, a speech signal with that acoustic feature being manipulated should be generated. We define

$$\bar{\mathbf{s}}_{\text{m}} = \bar{\mathbf{s}} + \frac{\hat{\mathbf{a}}}{||\hat{\mathbf{a}}||} \cdot \gamma \,, \tag{5}$$

where γ is a scaling factor for the manipulation and $\bar{\mathbf{s}}_{\text{m}}$ is the manipulated embedding, which is used as speaker input for the decoder.

Experiments

To train the FVAE, we employed two different data sets. The first is LibriVoxDeEn [5], a data set containing 86 German audio books from the open source platform LibriVox. It consists of ≈ 547 h of read German speech.

The goal of the FVAE is to disentangle content from style. So the task should be language independent, and therefore we took the LibriSpeech corpus as the second

Table 1: Disentanglement performance measured with PER on content embeddings (lower is better), EER on style embeddings (lower is better) and WER / CER on synthesized speech signal (lower is better). All values are stated in %.

#	Model	Training set	Test set	EER($\bar{\mathbf{s}}$)	PER($\mathbf{C}$)	WER($\hat{\mathbf{x}}$)	CER($\hat{\mathbf{x}}$)
1	Clean	-	NSC(de)	-	-	9.6	3.0
2	Feature extractor	-	NSC(de)	-	-	10.1	3.3
3	FVAE [1]	LibriSpeech(en)	LibriSpeech(en)	2.2	16.7	-	-
4	FVAE	LibriVoxDeEn(de)	NSC(de)	3.1	36.9	24.5	9.3
5	FVAE	LibriSpeech(en)	NSC(de)	2.0	34.5	22.2	8.4
6	FVAE	LibriSpeech(en) + LibriVoxDeEn(de)	NSC(de)	1.69	34.5	21.33	8.0

training corpus, despite containing the speech of English audiobooks. The subsets *train-clean-100* $\approx$ 100 h and *train-clean-360* $\approx$ 360 h are used for training, with 251 and 921 speakers each. Following [1], 60 % of the speaker utterances from *train-clean-100* are used for training.

The validation and testing is done on the Nautilus Speaker Characterization (NSC) corpus [6]. This is a German data set, which consists of $\approx$ 8 h scripted and $\approx$ 15 h of semi-spontaneous dialogues, with a total of 300 speakers. We used the semi-spontaneous dialogues for validation and the scripted dialogues for testing.

In all trainings, the utterances are segmented to a maximum length of 4 s and segments shorter than 2 s are discarded. To compensate for the mismatch between the data sets, all audio data is normalized to zero mean, a standard deviation (std) of 0.02 and resampled with a sampling frequency of 24 kHz. Training of the FVAE is performed with a batch size of 32 and a learning rate of $5 \cdot 10^{-4}$. The checkpoint that achieves the lowest reconstruction error on the validation set is used to report the results on the test set.

The performance of the FVAE is evaluated regarding the disentanglement of speaker and content. Following [1], we perform a speaker verification evaluation on the style embedding and report the Equal Error Rate (EER), where lower EER indicates better style embeddings. To measure the linguistic information of the content embedding, we trained a phone classifier on $\mathbf{C}$, similar to [1], but using 65 target classes. We used [7] to extract the phonemes. To assess the intelligibility of the synthesized speech $\hat{\mathbf{x}}$, we report the Word Error Rate (WER) and the Character Error Rate (CER) of Automatic Speech Recognition (ASR) experiments, which were conducted with a pretrained German recognizer [8].

German speech disentanglement

Table 1 shows the results for different training- and test set combinations. As a reference, row #1 reports the result of the ASR model trained on clean data and row #2 when a HiFi-GAN [9] is used to synthesize the time-domain signal from log-mel features. The latter WER marks the top line for the other models evaluated on the German data set.

Feature	Input $\rho(u, y)$	Reconstruction $\rho(u, \hat{y})$
Pitch	0.961	0.960
Hammarberg index	0.845	0.790

Table 2: Correlation between the canonical variable u and the extracted acoustic feature from the input y and output $\hat{y}$

Row #3 shows the disentanglement performance of the FVAE evaluated on LibriSpeech and the following two rows present the performance on the NSC data set. It can be seen that using LibriSpeech as training set, better performance is achieved, despite the language mismatch. But comparing these results with those achieved on the LibriSpeech test set, it is obvious, that the performance in terms of Phone Error Rate (PER) is much worse. Further, comparing the WERs on the NSC test set with the top line in row #2, it can be seen, that the model seems to generate too many artifacts, leading to an increase in WER. The last row shows the result of a model trained on LibriSpeech, where the training is continued on LibriVoxDeEn, while the encoders were frozen. The performance of the EER and the WER improved slightly. Note, that the test set was normalized with the combined training statistics of LibriSpeech and LibriVoxDeEn, while the model or row #5 only used the training statistics of LibriSpeech explaining the difference in EER despite the frozen encoders.

Style embedding manipulation

The speaker vectors of the model of row #6 of Table 1 are investigated in the following. Two projection vectors, one, $\hat{\mathbf{a}}_\mathrm{p}$, for the pitch and another, $\hat{\mathbf{a}}_\mathrm{h}$, for the Hammarberg index, are estimated on the German training set according to Eq. (4). The associated acoustic features y are determined on clean data and the speaker embeddings $\bar{\mathbf{s}}_\mathrm{m}$ are extracted from the speaker encoder. Fig. 3

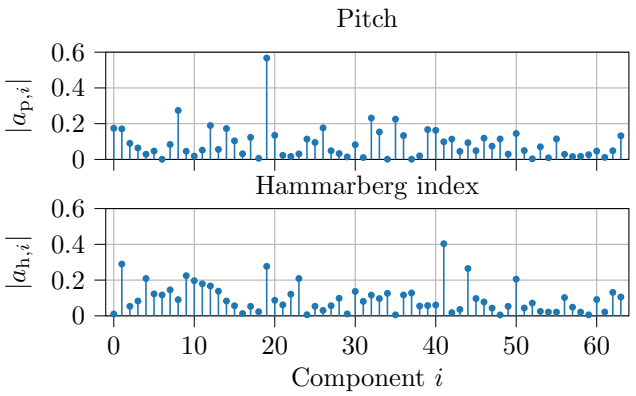

Figure 3: Absolute values of the projection vectors

shows the absolute values of the found projection vectors. The plot shows how strong each component of the speaker embedding contributes to the correlation with the acoustic feature. Obviously, and unfortunately, the information about the acoustic feature is widely spread over the components of the speaker vector.

To quantify, how much of the acoustic feature is encoded in the speaker embedding, the correlation between the canonical variable $u = \hat{\mathbf{a}}^{\mathrm{T}}\bar{\mathbf{s}}$ and the acoustic feature y extracted from the input signal is determined on the test set. The results are shown in the second column of Table 2. As indicated by the high correlation, there is a strong linear dependency between the projection of the speaker vector and the acoustic feature. To test whether this property also holds for the synthesized voice, the correlation between the canonical variables and the acoustic features extracted from the synthesized voice are calculated. These results are shown in the third column. It can be seen, that the correlation regarding the Hammarberg index is slightly reduced.

As a last experiment, the style embedding is manipulated according to Eq. (5), in an effort to modify the acoustic feature in the synthesized speech. The pitch of 30 male and female speakers with 5 random utterances each were manipulated. The same was also done with the Hammarberg index. Figure 4 shows the mean value of the extracted acoustic feature $\hat{y}$ of the synthesized voice, plotted over the manipulation factor γ, where $\gamma = 0$ denotes the reconstruction without manipulation. It can be seen, that the acoustic feature of the synthesized voice can be manipulated by shifting the style embedding, whereby the strength of manipulation is controlled by γ.

Conclusions

In this paper, we took a closer look at the speaker embedding vectors obtained from a FVAE that encodes speaker and content induced variations in different embedding vectors. Using CCA, we determined the directions in the latent speaker space that are strongest correlated with the mean average pitch and the Hammarberg index. We showed that these two acoustic features can be manipulated by shifting the speaker embeddings along those found directions. However, the acoustic properties are not well aligned with the coordinate axes of the latent speaker space. Thus, the modification of one acoustic feature will also affect the other, which limits the usefulness of acoustic feature modification though manipulation of the latent speaker vectors.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): TRR 318/1 2021 - 438445824.

References

[1] J. Ebbers, M. Kuhlmann, T. Cord-Landwehr, and R. Haeb-Umbach, "Contrastive predictive coding supported factorized variational autoencoder for unsupervised learning of disentangled speech representations," IEEE, 2021.

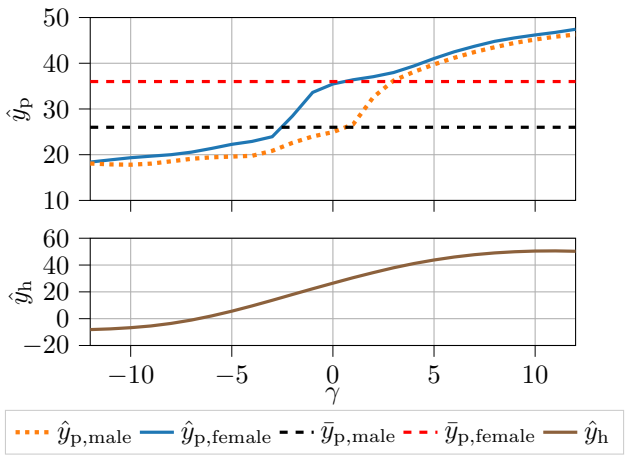

Figure 4: Extracted acoustic features of synthesized speech signals $\hat{y}$, where the style embeddings are manipulated in the direction of the highest correlation of an acoustic feature. γ defines the strength of manipulation and $\bar{y}_{\mathrm{p,male}}$ is the mean pitch of all male speakers in the NSC dataset, same for $\bar{y}_{\mathrm{p,female}}$. Pitch is measured in semitones.

[2] M. Kuhlmann, F. Seebauer, J. Ebbers, P. Wagner, and R. Haeb-Umbach, "Investigation into Target Speaking Rate Adaptation for Voice Conversion," in *Proc. Interspeech 2022*, pp. 4930–4934, 2022.

[3] 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*, pp. 1459–1462, 2010.

[4] L. Tamarit, M. Goudbeek, and K. Scherer, "Spectral slope measurements in emotionally expressive speech," *Proceedings of speech analysis and processing for knowledge discovery*, pp. 169–183, 2008.

[5] B. Beilharz, X. Sun, S. Karimova, and S. Riezler, "LibriVoxDeEn: A corpus for german-to-english speech translation and german speech recognition," *arXiv preprint arXiv:1910.07924*, 2019.

[6] L. F. Gallardo and B. Weiss, "The nautilus speaker characterization corpus: Speech recordings and labels of speaker characteristics and voice descriptions," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[7] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech*, pp. 498–502, 2017.

[8] H. N. Krabbenhöft and E. Barth, "TEVR: Improving Speech Recognition by Token Entropy Variance Reduction," *arXiv preprint arXiv:2206.12693*, 2022.

[9] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.