THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Disentangling style factors from speaker representations

Citation for published version:
Williams, J & King, S 2019, Disentangling style factors from speaker representations. in *Proceedings Interspeech 2019.* vol. 2019-September, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, ISCA, pp. 3945-3949, 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, Graz, Austria, 15/09/19. https://doi.org/10.21437/Interspeech.2019-1769

Digital Object Identifier (DOI):
10.21437/Interspeech.2019-1769

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings Interspeech 2019

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 24. Apr. 2024

Disentangling Style Factors from Speaker Representations

Jennifer Williams and Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
j.williams@ed.ac.uk

Abstract

Our goal is to separate out speaking style from speaker identity in utterance-level representations of speech such as *i*-vectors and *x*-vectors. We first show that both *i*-vectors and *x*-vectors contain information not only about speaker but also about speaking style (for one data set) or emotion (for another data set), even when projected into a low-dimensional space. To disentangle these factors, we use an autoencoder in which the latent space is split into two subspaces. The entangled information about speaker and style/emotion is pushed apart by the use of auxiliary classifiers that take one of the two latent subspaces as input and that are jointly learned with the autoencoder. We evaluate how well the latent subspaces separate the factors by using them as input to separate style/emotion classification tasks. In traditional speaker identification tasks, speaker-invariant characteristics are factorized from channel and then the channel information is ignored. Our results suggest that this so-called channel may contain exploitable information, which we refer to as *style factors*. Finally, we propose future work to use information theory to formalize *style factors* in the context of speaker identity.

Index Terms: speaking style, emotion recognition, speech disentanglement, speaker recognition

1. Introduction

Speaker embeddings, such as *i*-vectors and *x*-vectors, that have been originally designed to perform well in speaker identification tasks also contain information about speaking style and emotion. They are optimized to model speaker identity for tasks such as speaker recognition, speaker verification, and speaker diarization. These vectors maximize the variance between speakers, while minimizing within-speaker differences (recording device, mood, age, etc) [1]. We show that style and emotion information can be found within those same utterance-level speaker embeddings. Our ultimate goal is to develop a framework that can separate style, channel and speaker information because different factors of the speech signal could then be exploited for different uses, for example in speech synthesis.

In the process of training high-quality speaker embeddings, one necessary step is to separate speaker-invariant characteristics from residual channel information. The channel information contains factors related to the recording device and session noise and is removed to address differences such as the mismatch between microphone and low-quality telephony speech. The final speaker embeddings thus discard channel information and retain speaker-invariant characteristics. In both *i*-vectors and *x*-vectors, the channel factorization is performed during Probabilistic Linear Discriminant Analysis (PLDA). In this work, we therefore utilize the utterance-level representations obtained before PLDA.

This paper adopts a working definition of style to be: *how speakers adapt their speaking manner according to the speaking context*. In this work we investigate four categories of speaking style: spontaneous conversation, goal-directed interaction, retold passage, and read passage. In separate experiments on other data, we also explore four basic categories of emotion: angry, happy, sad, and neutral. This paper makes three main contributions. First, to show that utterance-level *i*-vectors and *x*-vectors contain information about speaking style and emotion. Second, to compare disentanglement methods, evaluated using accuracy on a classification task. Finally, we demonstrate that both style and residual information are necessary components of utterance-level speaker embeddings in order to minimize reconstruction error.

There is some evidence that *i*-vectors discriminate emotion in speech [2]. We believe the information is found within the channel factor; therefore *x*-vectors will also contain it. It consists of structure and content that correlates to meaningful categories of style and emotion, which we refer to as *style factors*.

2. Related Work

2.1. Speaking Style

There are systematic and measurable differences when a speaker expresses a particular emotion or speaks in a particular context. Two well-studied contexts include *read* and *spontaneous* speech. For example, differences were found between news broadcast speech and freestyle conversations and these differences include intonation patterns ($F0$) and speaking rate [3]. Previous work has shown that content words are more likely to carry marked pitch accents in spontaneous speech than in read speech [4]. On the other hand, there are fewer pitch accents overall in spontaneous speech and dialect does not seem to influence variation as much as the speaking style [5].

Importantly, speaker-dependent style modeling does not appear to improve pitch accent modeling and a more general approach to style modeling is recommended [4]. Work from [6] showed that sequential modeling of symbolic prosodic annotations (TOBI) can differentiate between four different speaking styles better than direct modeling of signal-level features (such as $F0$). The four speaking styles included: broadcast news, spontaneous monologues, read monologues, and interactive dialogue. [7] first showed that GMM supervectors can be used to classify consecutive utterances of spontaneous and read speech with $> 95\%$ accuracy, independently of any knowledge about speaker identity. Their use of GMM supervectors aligns to our work in this paper because GMM supervectors form the basis for *i*-vectors, one of the embedding types we explore. All this prior work suggests it is possible to arrive at a canonical representation of speaking style, and has influenced the approach we take here. Our work extends to *x*-vectors, and we also explore additional categories of style and emotion.

2.2. Expressive Speech Synthesis

Expressive speech synthesis is one application that benefits from robust representations of style and emotion. Recent work in this area has explored using textual cues for determining paralinguistic style [8, 9] as well as using held-out labeled data to drive controllability [10]. Recently [9] introduced a method to construct and apply *style tokens* in an effort to model style and speaking rate. These tokens, which are essentially learned embeddings, represent a set of discovered latent styles in speech data rather than human-labeled categories. The final style embeddings are not convincingly categorical to listeners. Further, as in [11], there is an underlying assumption that *style tokens* can be learned directly from features that have been extracted from the speech signal, such as $F0$, and this is somewhat contradictory to other work that calls for higher-level abstract representations.

Others have explored abstract representations for expressive speech synthesis, including [2] who showed that *i*-vectors can form unsupervised clusters corresponding to emotion categories. They showed that *i*-vectors can be utilized for expressive synthesis. However, this approach did not remove speaker-invariant characteristics, making it difficult to control both speaker identity and emotion at the same time. Autoregressive models for expressive speech synthesis were explored in [12] using variational autoencoders (VAEs) with VoiceLoop, but they were unable to model global characteristics of style, possibly due to the approach being unsupervised.

2.3. Speaker Identification

This paper explores two types of utterance-level speaker embeddings from the field of speaker identification: *i*-vectors [13] and *x*-vectors [14]. Both are considered the state-of-the-art for automatic speaker recognition and continue to be fine-tuned and adapted for speech processing tasks [15, 16, 17, 18]. *x*-vectors claim improvement over *i*-vectors in part because they do not rely on content information during training [19]. Both of these speaker embeddings are evaluated with Probabilistic Latent Discriminant Analysis (PLDA) [20]. That is where channel and session residual information is separated from speaker-invariant characteristics.

3. Datasets

3.1. IViE

The Intonational Variations in English (IViE) corpus was originally collected for the purpose of exploring 9 regional/dialect variations throughout the United Kingdom [21]. From the set of parallel tasks, we have selected four style categories: spontaneous conversation, goal-directed interaction, retold passages, and read passages. Each of the speakers is approximately 16 years old, and the collection was split evenly between males and females. We disregarded dialect and gender labels, instead focusing only on these four style categories. For the spontaneous conversation and goal-directed interaction categories the audio was not diarized so we combined both speaker labels together and set the label as the first speaker. The spontaneous conversation category used same-gender pairs who discussed the topic of cigarette smoking. The goal-directed interaction involved a version of the conventional map-task (i.e. providing directions while reading a map). The retold and read passages involved an excerpt from the story *Cinderella*[1]. A description of the data split, number of unique speakers, and average utterance duration (in seconds) is provided in Table 1.

Table 1: *Number of utterances for IEMOCAP and IViE datasets, where S is the number of unique speakers in the category, and dur is the average duration of audio segments in seconds.*

		Train	**Valid**	**Test**	S	$dur(s)$
IEMOCAP	**Angry**	318	56	125	10	4.8
	Sad	269	48	106	10	5.6
	Happy	60	10	23	10	5.0
	Neutral	247	44	97	10	4.4
	Totals	894	158	351	10	4.9
IViE	**Conv.**	210	37	82	108	50
	Directed	426	75	168	112	31
	Retold	558	99	219	110	27
	Read	431	76	169	110	44
	Totals	1625	287	638	112	38

3.2. IEMOCAP

We also used the IEMOCAP dataset [22] for emotion categories. 10 professional actors were prompted to enact hypothetical situations or to read directly from a script while performing emotions. Although this dataset is multimodal (speech, video, head movements, transcription, etc), we only use the audio. The data annotation and labeling procedure allowed utterances to have multiple labels, so we identified a subset of four emotions wherein the utterances had only one label. Our subset contains: anger, sadness, happiness and neutral. More details about the data split is provided in Table 1. While this significantly reduced the overall size of our data, it circumvented issues of label reliability from the annotators. We used the IEMOCAP dataset to make a comparison to previous works in our separate classification task. Recently [23] achieved 68.8% overall classification accuracy working directly on the speech spectrograms. We did not have information about their train/test split, so although our classifier exceeded this baseline it could be due to differences in training sets and labels.

4. Methodology

We project utterance-level embeddings into a pair of latent spaces. One is intended to contain *only* style (or emotion) information, and the other to contain *no* style (or emotion) information. To quantify how disentangled the latent spaces are, we use them as input to a style (or emotion) classifier, and examine how high or low the accuracy is. To further demonstrate that the two latent spaces contain complementary information, we perform reconstruction experiments wherein the latent space containing *no* style (or emotion) has been degraded. If one latent space has been degraded and the autoencoder is unable to reconstruct the input, then we know the two latent spaces contain complimentary information. Our code can be found at `https://github.com/rhoposit/style_factors`

4.1. Utterance-Level Embeddings

We used the Kaldi Toolkit [24] to extract utterance-level *i*-vectors and *x*-vectors. We employed the pre-trained models[2] described in [14], which provided us with PLDA, mean vectors, and transform vectors which had been trained and evaluated on the VoxCeleb corpus. That corpus contains approximately 2,000 unique celebrity speakers [25, 26] and was augmented with noise during training. The VoxCeleb data is considered

[1] `http://www.phon.ox.ac.uk/files/apps/IViE/stimuli.php`

[2] `http://kaldi-asr.org/models/m7`

spontaneous and 'in the wild' and it is also known to exhibit natural emotion [27]. The front-end configuration was provided to us fully-specified with the following settings. The audio signal sampling rate was 16 kHz and the frame length was 25 ms. For feature-level vocal-tract length normalization (VTLN), the the low-frequency cutoff was 20 Hz and the high-frequency cutoff was 7600 Hz. The features were 24 MFCCs for *i*-vectors and 30 MFCCs for *x*-vectors. The number of mel-cepstrum filterbank bins was 30. The features were mean-normalized over a sliding window up to 3 seconds.

The internal contents of these utterance-level embeddings come from DNNs. For *x*-vectors it is the first *segment* layer after frame-level statistics pooling, where each utterance-level vector is 1x512. The non-linearities during training come from ReLU activation. The *i*-vectors come from DNN phonetic bottleneck features of a time-delay neural network (TDNN) with p-norm activations where each utterance-level vector is 1x400 [19].

4.2. Dimensionality Reduction

As a point of comparison for the later disentangled latent spaces, we wanted to know to what extent the *i*-vector and *x*-vector utterance embeddings could be compressed while retaining style (or emotion) information. Using either PCA or the vanilla autoencoder of Figure 1, we projected the embeddings into latent spaces of dimension $dims = [512, 400, 300, 200, 100, 50, 20, 10]$, omitting 512-*dim* when using *i*-vectors. All such autoencoders consisted of four fully-connected dense layers for the encoder and four fully-connected dense layers for the decoder with the same training parameters as our DNN classifier: ReLU activation [28], $L2$ regularization [29], Adam optimizer [30] with learning rate $lr = 0.0002$, and early-stopping monitored by validation loss. The input was normalized.

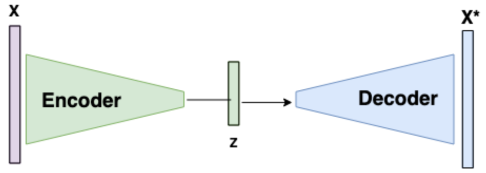

Figure 1: *Vanilla autoencoder (AEV)*

4.3. Disentanglement

We propose the autoencoder with two encoders and two auxiliary classifiers shown in Figure 2. The $z1$ and $z2$ latent spaces have separate auxiliary classifiers. To cause $z1$ to encode style, its auxiliary style classifier is trained to minimize cross-entropy loss. First we established a baseline autoencoder using two latent spaces (AE1) and similar auxiliary classifiers, shown in Figure 2. To cause the residual latent space $z2$ to contain as *little* style information as possible, we experimented with variants including resetting the $z2$ space to the batch mean (AE2), maximizing the auxiliary classifier cross-entropy loss (AE3), or completely degrading the $z2$ space and re-training the decoder (AEC). When we used maximized cross-entropy loss on the residual $z2$ auxiliary classifier, we adjust that task loss weight to $w = 0.05$ as this value helped balance the weight of all the losses, though we acknowledge that setting optimal loss weights is itself an open problem in multi-task learning. A description of the set of autoencoder configurations is presented in Table 2. It is worth noting that we had explored additional autoencoders, for example with and without a split latent space, but for this paper we selected those which had highest $z1$ classification and demonstrate $z1$ and $z2$ are both needed for reconstruction.

Table 2: *Description of disentanglement methods. PCA and AEV do not attempt disentanglement are used for comparison.*

Technique	Num Encoders	$z2$ Corruption	Max Loss
PCA	–	–	no
AEV	1	–	no
AE1	2	–	no
AE2	2	μ batch	no
AE3	2	–	yes
AEC	2	full	yes

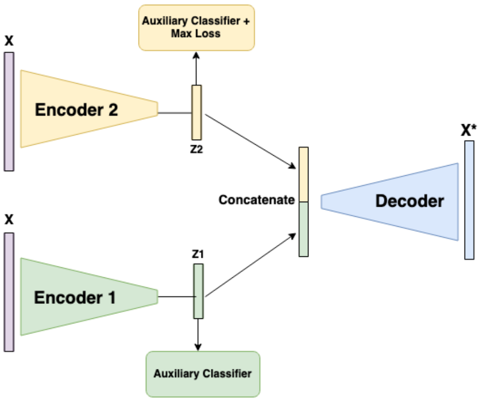

Figure 2: *Autoencoder with dual-encoders and single decoder*

4.4. Style / Emotion Classification

Entirely distinct from the autoencoders, we trained DNN style (or emotion) classifiers using the Keras library [31] with TensorFlow backend [32]. We used the same train/test split as described in Table 1. This DNN consisted of three fully-connected dense layers with ReLU activation ($alpha = 0.2$) and $L2$ regularization ($L2 = 0.0001$). The optimizer was Adam with learning rate set to $lr = 0.0002$ and the remaining parameters were kept as default. The loss function was cross-entropy with early stopping. The input was normalized. We first trained the classifier to use raw *i*-vectors or *x*-vectors. Later we used the compressed representations from PCA or the vanilla autoencoder. Finally we used the disentangled latent spaces $z1$ or $z2$ provided by the dual autoencoder. In each case, a style (or emotion) classifier was trained from scratch.

5. Experiments and Results

5.1. Autoencoder Reconstruction Loss

One demonstration of disentanglement is to examine the ability to reconstruct the original input *i*-vector or *x*-vector from the latent space. As a reconstruction baseline, we calculated an average *i*-vector or *x*-vector over the training data and compared it to each training example to calculate the upper bound mean absolute error (MAE). Any MAE values above this bound indicate that the autoencoder reconstruction is very poor. The upper-bound baseline MAE for *i*-vectors was 0.75 for both IEMOCAP

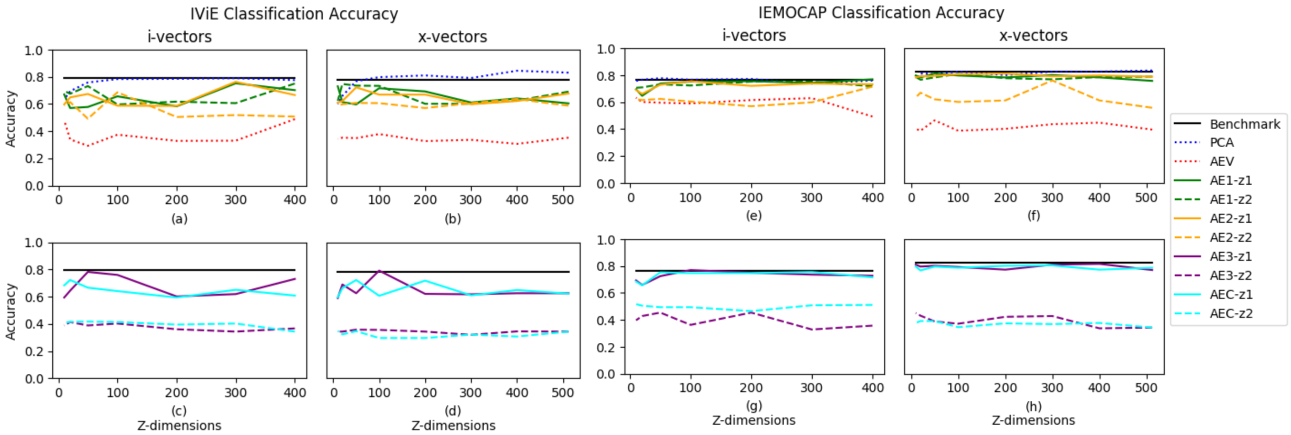

Figure 3: *Classification accuracy results before disentanglement (top) and after (bottom), with benchmarks constant for comparison. The benchmarks use raw i-vectors or x-vectors respectively as input and are shown in the plots as a constant horizontal line indicating classification accuracy without any compression or disentanglement. On IViE: 79% and 78%. For IEMOCAP: 76% and 82%.*

and IViE datasets. For x-vectors it was 0.62 for IEMOCAP and 0.63 for IViE. The MAE upper bound was exceeded with x-vectors when using the AEC model on the IViE data (MAE > 0.64). The MAE upper bound was approached and exceeded for the IEMOCAP data for both types of input vectors. High reconstruction error when $z2$ is corrupt indicates that *both* $z1$ and $z2$ components are needed. That is, they contain complementary – or disentangled – information.

5.2. Style / Emotion Classification Results

Our other demonstration of disentanglement was to use each of $z1$ and $z2$ in turn as the input to style (or emotion) classifiers described in Section 4.4. We report results for classification on held-out data in Figure 3. The upper set of plots in (a, b, e, f in Figure 3) show classification accuracy for methods that do not uniquely disentangle style in the $z1$ and $z2$ latent spaces. The bottom set of plots (c, d, g, h Figure 3) show classification accuracy for methods that were successful in isolating factors of style in $z1$ from residual in $z2$. Overall, the $z2$ space has lost information about style and emotion. On the other hand, the $z1$ space has preserved it through the range of latent dimensions, while continuing to perform close to benchmark. The overall best classification accuracy came from AE3.

In Figure 4 we show the confusion matrix for each dataset. These results are for the best-performing disentangled $z1$ encodings from AE3. The encoding size for IViE i-vectors was 50-*dim*, and for x-vectors was 100-*dim*. In the IViE style prediction tasks, we noticed that spontaneous conversation was often mistaken as retold speech. This may be a consequence of using non-diarized conversational speech although it did not seem to affect the goal-directed speech style which was also not diarized. The encoding size for IEMOCAP i-vectors was 300-*dim*, and x-vectors was 400-*dim*. The poor performance on 'happy' is related to class imbalance, similar to previous work [10] and is often identified as 'sad' or 'neutral'.

Further we compared how well the style factors in $z1$ and the residual in $z2$ retained *speaker* identifying information in an ad-hoc speaker PLDA evaluation. The original extracted i-vectors and x-vectors for IViE data discriminated speakers with below 10% equal-error-rate (EER) while for IEMOCAP the EER was above 30% EER. For all of the $z1$ and $z2$ representations in both datasets, the EER was always greater than 30%. This suggests speaker information was lost while style information was preserved.

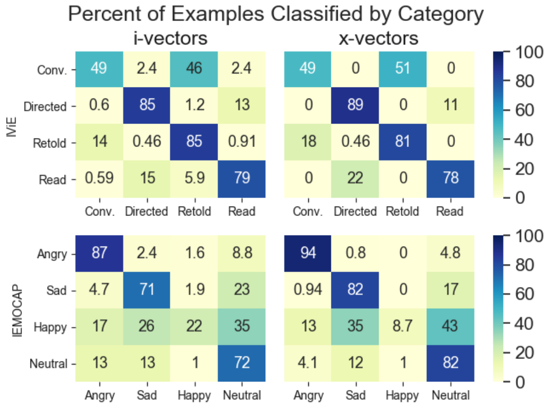

Figure 4: *Confusion matrices for per-category classification with IViE (top) and IEMOCAP (bottom).*

6. Discussion

We have demonstrated that two types of utterance-level representation invented for speaker identification, i-vectors and x-vectors, contain information that is predictive of style and emotion. This finding suggests the existence of *style factors* that are separate from channel and other speaker-invariant characteristics. Disentangling such factors would be highly useful in many speech applications including speech-to-speech translation, speech synthesis, and speaker identification.

7. Acknowledgements

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. The authors thank Steve Renals (CSTR), Junichi Yamagishi (NII, CSTR), Antreas Antoniou (IANC), and Aghilas Sini (IRISA) for helpful discussions.

8. References

[1] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] I. Jauk, "Unsupervised Learning for Expressive Speech Synthesis," *PhD Thesis. Universitat Politècnica de Catalunya*, 2017.

[3] J. Hirschberg, "A corpus-based approach to the study of speaking style," in *Prosody: Theory and Experiment*, pp. 335–350, Springer, 2000.

[4] J. Yuan, J. M. Brenier, and D. Jurafsky, "Pitch Accent Prediction: Effects of Genre and Speaker," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[5] H. Mixdorff, H. R. Pfitzinger, and K. Grauwinkel, "Towards Objective Measures for Comparing Speaking Styles," *ProcSPECOM*, pp. 131–134, 2005.

[6] A. Rosenberg, "Symbolic and Direct Sequential Modeling of Prosody for Classification of Speaking-Style and Nativeness," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[7] T. Asami, R. Masumura, H. Masataki, and S. Sakauchi, "Read and Spontaneous Speech Classification Based on Variance of GMM Supervectors," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting Expressive Speaking Style From Text In End-To-End Speech Synthesis," *arXiv preprint arXiv:1808.01410*, 2018.

[9] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-To-End Speech Synthesis," *arXiv preprint arXiv:1803.09017*, 2018.

[10] Z. Hodari, O. Watts, S. Ronanki, and S. King, "Learning Interpretable Control Dimensions for Speech Synthesis by Using External Data," *Proc. Interspeech 2018*, pp. 32–36, 2018.

[11] Y. Wang, R. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Saurous, "Uncovering Latent Style Factors for Expressive Speech Synthesis," *arXiv preprint arXiv:1711.00520*, 2017.

[12] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder," *arXiv preprint arXiv:1804.02135*, 2018.

[13] 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.

[14] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, IEEE, 2018.

[15] H. Zeinali, L. Burget, J. Rohdin, T. Stafylakis, and J. Cernocky, "How to Improve Your Speaker Embeddings Extractor in Generic Toolkits," *arXiv preprint arXiv:1811.02066*, 2018.

[16] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[17] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-Vector Length Normalization in Speaker Recognition Systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[18] M. McLaren, D. Castan, M. K. Nandwana, L. Ferrer, and E. Yilmaz, "How to Train Your Speaker Embeddings Extractor," *Les Sables d'Olonne, France: ISCA*, 2018.

[19] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," *Proc. Interspeech*, pp. 999–1003, 2017.

[20] S. J. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences About Identity," in *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, IEEE, 2007.

[21] E. Grabe, B. Post, and F. Nolan, "The IViE Corpus," *http://www.phon.ox.ac.uk/IViE*, 2011.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[23] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition From Speech Using Deep Learning on Spectrograms," in *Eighteenth Annual Conference of the International Speech Communication Association*, 2017.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, Dec. 2011. IEEE Catalog No.: CFP11SRW-USB.

[25] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A Large-Scale Speaker Identification Dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[26] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep Speaker Recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[27] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion Recognition in Speech using Cross-Modal Transfer in the Wild," in *ACM Multimedia*, 2018.

[28] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital Selection and Analogue Amplification Coexist in a Cortex-Inspired Silicon Circuit," *Nature*, vol. 405, no. 6789, p. 947, 2000.

[29] A. Y. Ng, "Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance," in *Proceedings of the Twenty-First International Conference on Machine learning*, p. 78, ACM, 2004.

[30] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A System for Large-Scale Machine Learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 265–283, 2016.