DEEP FACTORIZATION FOR SPEECH SIGNAL

Lantian Li, Dong Wang, Yixiang Chen, Ying Shi, Zhiyuan Tang, Thomas Fang Zheng

Center for Speech and Language Technologies, Research Institute of Information Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

ABSTRACT

Various informative factors mixed in speech signals, leading to great difficulty when decoding any of the factors. An intuitive idea is to factorize each speech frame into individual informative factors, though it turns out to be highly difficult. Recently, we found that speaker traits, which were assumed to be long-term distributional properties, are actually short-time patterns, and can be learned by a carefully designed deep neural network (DNN). This discovery motivated a *cascade deep factorization* (CDF) framework that will be presented in this paper. The proposed framework infers speech factors in a sequential way, where factors previously inferred are used as conditional variables when inferring other factors. We will show that this approach can effectively factorize speech signals, and using these factors, the original speech spectrum can be recovered with a high accuracy. This factorization and reconstruction approach provides potential values for many speech processing tasks, e.g., speaker recognition and emotion recognition, as will be demonstrated in the paper.

Index Terms— speech signal processing, speech recognition, speaker recognition, emotion recognition

1. INTRODUCTION

Speech signals involve rich information, including linguistic content, speaker trait, emotion, channel and background noise, etc. Researchers have worked for several decades to decode these information, leading to a multitude of speech information processing tasks, including automatic speech recognition (ASR) and speaker recognition (SRE) [1]. After a long-term research, some tasks have been addressed pretty well, at least when a large amount of data is available, e.g., ASR and SRE; while others remain difficult, e.g., automatic emotion recognition (AER) [2].

A major difficulty of speech processing resides in the fact that multiple informative factors are intermingled together, and so whenever we decode for a particular factor, all other factors contribute as uncertainties. An intuitive idea to deal with the information blending is to factorize the speech signal into individual informative factors at the frame level. However, it turns out to be highly difficult, due to at least two reasons: Firstly, the way that these factors are mixed is unclear and seems highly complex; Secondly, and perhaps more fundamentally, some major factors, particularly the speaker trait, behaves as long-term distributional properties rather than short-time patterns. It has been partly demonstrated by the fact that most of the successful speaker recognition approaches (e.g., JFA [3], i-vector [4]) rely on statistical models that retrieve speaker vectors based multiple frames (segments). Therefore, there is a wide suspicion that speech signals are short-time factorizable.

Fortunately, our recent study showed that speaker traits are largely short-time spectral patterns, and a carefully designed deep neural network can learn to extract these patterns at the frame level [5]. The following studies demonstrated that the frame-level deep speaker features are highly generalizable: they work well with voices of trivial events, such as laugh and cough that are as short as 0.3 seconds [6]; and they are robust against language mismatch [7]. The short-time property of speaker traits suggests that speech signals are possibly short-time factorizable, as it has been known that another major speech factor, the linguistic content, is also short-time identifiable [8].

In this paper, we present a *cascaded deep factorization* (CDF) approach to obtain such factorization. By this approach, the most significant factors are inferred firstly, and other less significant factors are inferred subsequently on the condition of the factors that have already been inferred. Our experiments on a speaker recognition task and an emotion recognition task demonstrated that the CDF-based factorization is highly effective. Furthermore, we show that the original speech signal can be reconstructed from the CDF-derived factors pretty well.

2. SPEAKER FEATURE LEARNING

In the previous study [5], we presented a CT-DNN structure that can learn speaker features at the frame level. The network consists of a convolutional (CN) component and a time-delay (TD) component. The CN component comprises two CN layers, each followed by a max-pooling layer. The TD component comprises two TD layers, each followed by a P-norm layer. The output of the second P-norm layer is projected into a feature layer. The activations of the units of this layer, after length normalization, form the speaker feature of the input speech frame. During model training, the feature layer is fully connected to an output layer whose units correspond to the speaker identities in the training data. The train-

This work was supported by the National Natural Science Foundation of China under Grant No.61633013 / 61371136 and the National Basic Research Program (973 Program) of China under Grant No.2013CB329302. A pre-print version was published on arXiv:1706.01777. Dong Wang is the corresponding author (wangdong99@mails.tsinghua.edu.cn).

ing is performed to optimize the cross-entropy objective that aims to discriminate the training speakers based on the input frames. We demonstrated that the speaker feature inferred by the CT-DNN structure is highly speaker-discriminative [5], and speaker traits are largely short-time spectral patterns and can be identified at the frame level.

3. CASCADED DEEP FACTORIZATION

The ASR research has demonstrated that the linguistic content can be individually inferred by a DNN at the frame level [9], and the deep speaker feature learning method described in the previous section demonstrated that speaker traits can be also identified by a very short segment. We denote this single factor inference method based on deep neural models by *individual deep factorization* (IDF). The rationality of the IDF method is two-fold: Firstly, the target factor (linguistic content or speaker trait) is sufficiently significant in speech signals; Secondly, a large amount of training data is available. It is the large-scale supervised learning that picks up the most task-relevant factors by leveraging the power of DNNs in feature learning. For factors that are less significant and/or without much training data, however, IDF is not applicable. Fortunately, the successful inference of the linguistic and the speaker factors may significantly simplify the inference of those 'not so prominent' factors. This motivated a cascaded deep factorization (CDF) approach: firstly we infer a particular factor by IDF, and then use this factor as a conditional variable to infer the second factor, and so on. Finally, the speech signal will be factorized into a set of individual factors, each corresponding to a particular task.

To demonstrate this concept, we apply the CDF approach to factorize emotional speech signals into three factors: linguistic, speaker and emotion. Fig. 1 illustrates the architecture. Firstly an ASR system is trained using word-labelled speech data. The frame-level linguistic factor, which is in the form of phone posteriors in our study, is produced from the ASR system, and is concatenated with the raw feature (Fbanks) to train an SRE system. This SRE system is used to produce the frame-level speaker factor, as discussed in the previous section. The linguistic factor and the speaker factor are finally concatenated with the raw feature to train an AER system, by which the emotion factor is produced from the last hidden layer.

The CDF approach is fundamentally different from the conventional factorization approach, e.g., JFA [3]: (1) CDF is frame-level while conventional methods are segment-level; (2) CDF relies on discriminative training while conventional factorization methods rely on maximum likelihood estimation; (3) CDF infers factors sequentially and so can use data with partial labels (e.g., only speaker labels), while conventional approaches infer factors jointly so can only use full-labelled data; (4) CDF are based on DNNs that are deep, non-linear and non-Gaussian, while most conventional approaches are based on models that are shallow, linear and Gaussian.

We highlight that more complex model structures are possible to conduct the factorization, e.g., with the collaborative learning architecture [10]. However, the CDF framework is consistent with a cascaded convolution view for speech signals, i.e., speech signals are produced by convolving informative factors sequentially, from linguistic parts to non-linguistic parts, as mentioned in [11, 12].

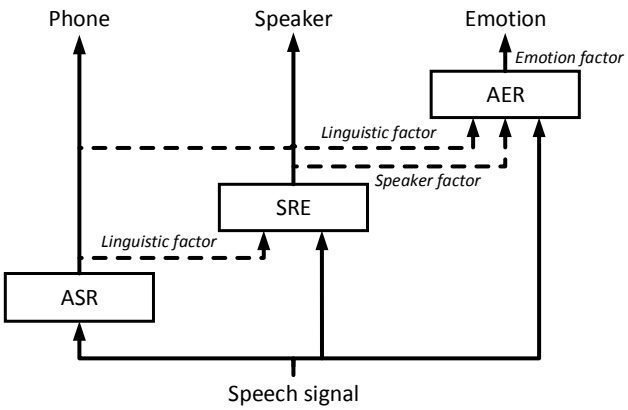

Fig. 1. The cascaded deep factorization approach applied to factorize emotional speech into three factors: linguistic, speaker and emotion.

4. SPECTRUM RECONSTRUCTION

An interesting property of the CDF-inferred factors is that they can be used to recover the original speech. Define the linguistic factor q, the speaker factor s, and the emotion factor e. For each speech frame, we try to use these three factors to recover the spectrum x. According to the cascaded convolution view, the reconstruction is written in the form:

$$ln(x) = ln\{f(q)\} + ln\{g(s)\} + ln\{h(e)\} + \epsilon \qquad (1)$$

where f, g, h are the non-linear recovery function for q, s and e respectively, each implemented as a DNN. ϵ represents the residual which is assumed to be Gaussian. This reconstruction is illustrated in Fig. 2, where all the spectra are in the log domain. Note that q, s, e are all inferred from Fbanks rather than the original spectra, but they can still recover the original signal pretty well, as will be seen in Section 6.

5. RELATED WORK

The CDF approach shared a similar motivation as the phonetic DNN i-vector approach [13, 14]: both utilize the phonetic factor to support inference of other factors. The difference is that CDF is a neural model and retrieves frame-level features, while phonetic DNN i-vector is a probabilistic model and retrieves utterance-level representations. CDF is also related to multi-task learning [15] and transfer learning [16, 17], where multiple tasks are used to regularize the training [10, 18, 19]. Compared to these methods, the CDF approach focuses on explaining variabilities of speech signals, rather than learning models. Finally, the CDF approach is also related to auto-encoder (AE) that can be regarded as an unsupervised factor-

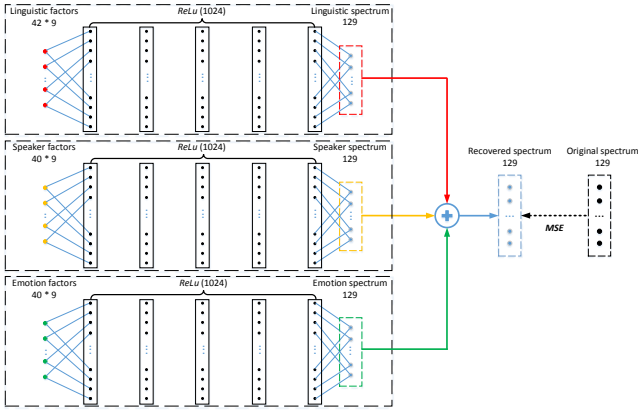

Fig. 2. The architecture for cascaded convolutional spectrum reconstruction.

ization. Compared to AE, CDF is a supervised learning, and the learned factors are inherently task-oriented.

6. EXPERIMENT

6.1. Database

Three databases are used in our experiment, as presented below. All the speech signals are down-sampled to 8k Hz to ensure data consistency.

ASR database: The *WSJ* database was used to train the ASR system. The training data is the official *train_si284* dataset, composed of 282 speakers and $37,318$ utterances. The test set contains three datasets (*devl92, eval92 and eval93*), including 27 speakers and $1,049$ utterances in total.

SRE database: The *Fisher* database was used to train the SRE systems. The training set consists of $2,500$ male and $2,500$ female speakers, with $95,167$ utterances in total. Each speaker has about 120 seconds of speech signals. It was used for training the UBM, T-matrix and PLDA models of an i-vector baseline system, and the DNN model described in Section 2. The evaluation set consists of 500 male and 500 female speakers, $82,990$ utterances in total. The speakers of the training set and the evaluation set are not overlapped. For each speaker, 10 utterances (about 30 seconds in total) are used for enrollment and the rest for test.

AER database: The *CHEAVD* database [20] was used to train the AER systems. This database was selected from Chinese movies and TV programs and was used as the standard database for the multimodal emotion recognition challenge (MEC 2016) [21]. There are 8 emotions in total: Happy, Angry, Surprise, Disgust, Neutral, Worried, Anxious and Sad. The training set contains $2,224$ utterances and the evaluation set contains 628 utterances.

6.2. ASR baseline

We first build a DNN-based ASR system using the WSJ database. This system will be used to produce the linguistic factor in the following CDF experiments. The Kaldi toolkit [22] is used to train the DNN model, following the Kaldi WSJ s5 nnet recipe. The DNN structure consists of 4 hidden layers, each containing $1,024$ units. The input feature is Fbanks, and the output layer discriminates $3,383$ GMM pdfs. With the official 3-gram language model, the word error rate (WER) of this system is 9.16%. The linguistic factor is represented by the 42-dimensional phone posteriors, derived from the output of the ASR DNN.

6.3. SRE by IDF and CDF

We build three SRE systems: an i-vector/PLDA system [4, 23] to represent the conventional statistical model approach; an IDF d-vector system that follows the CT-DNN architecture, where only the raw features (Fbanks) comprise the input; and a CDF d-vector system that follows the CDF spirit and the linguistic factors produced by the ASR system are used as additional input to the CT-DNN architecture.

For the i-vector system, the UBM is composed of $2,048$ Gaussian components, and the i-vector dimension is set to 400. The system is trained following the Kaldi SRE08 recipe. For the d-vector systems, the frame-level speaker features are of 40 dimensions, and the utterance-level d-vector is derived as an average of the frame-level features within the utterance. More details of the d-vector systems can be found in [5].

We report the results on the identification task, though similar observations were obtained on the verification task. In the identification task, one matched speaker (*Top-1*) is identified given a test utterance. With the i-vector (d-vector) system, each enrolled speaker is represented by the i-vector (d-vector) of their enrolled speech, and the i-vector (d-vector) of the test speech is compared with the i-vectors (d-vectors) of the enrolled speakers, finding the speaker whose enrollment i-vector (d-vector) is nearest to that of the test speech. For the i-vector system, the popular PLDA model [23] is used to measure the similarity between i-vectors; for the d-vector system, the simple cosine distance is used.

The results in terms of the Top-1 identification rate (IDR) are shown in Table 1. In this table, 'C(30-20f)' means the test condition where the duration of the enrollment speech is 30 seconds, while the test speech is 20 frames. Note that 20 frames is just the length of the effective context window of the speaker CT-DNN, so only one single frame of speaker feature is used in this condition. From these results, it can be observed that the d-vector system performs much better than the i-vector baseline, particularly with very short speech segments. Comparing the IDF and CDF results, it can be seen that the CDF approach that involves phonetic knowledge as the conditional variable greatly improves the d-vector system in the short speech segment condition.

Table 1. The ***Top-1*** IDR(%) results on the short-time speaker identification with the i-vector and two d-vector systems.

		IDR%		
Systems	Metric	C(30-20f)	C(30-50f)	C(30-100f)
i-vector	PLDA	5.72	27.77	55.06
d-vector (IDF)	Cosine	37.18	51.24	**65.31**
d-vector (CDF)	Cosine	**47.63**	**57.72**	64.45

6.4. AER by CDF

This section applies the CDF approach to an emotion recognition task. For that purpose, we first build a DNN-based AER baseline. The DNN model consists of 6 hidden layers, each containing 200 units. After each layer, a P-norm layer reduces the dimensionality from 200 to 40. The output layer comprises 8 units, corresponding to the number of emotion classes in the CHEAVD database. This DNN model produces frame-level emotion posteriors. The utterance-level posteriors are obtained by averaging the frame-level posteriors, by which the utterance-level emotion decision is achieved.

Three CDF configurations are investigated, according to which factor is used as the conditional: the linguistic factor (+ ling.), the speaker factor (+ spk.) and both (+ ling. & spk.). The results are evaluated in two metrics: the identification accuracy (ACC) that is the ratio of the correct identification on all emotion categories; the macro average precision (MAP) that is the average of the ACC on each of the emotion category.

Table 2. Accuracy (ACC) and macro average precision (MAP) of the AER systems.

	Training set			
	ACC% (fr.)	MAP% (fr.)	ACC% (utt.)	MAP% (utt.)
Baseline	74.19	61.67	92.27	83.08
+ling.	86.34	81.47	96.94	96.63
+spk.	92.56	90.55	97.75	97.16
+ling. & spk.	**94.59**	**92.98**	**98.02**	**97.34**
	Evaluation set			
	ACC% (fr.)	MAP% (fr.)	ACC% (utt.)	MAP% (utt.)
Baseline	23.39	21.08	28.98	24.95
+ling.	27.25	27.68	**33.12**	**33.28**
+spk.	27.18	28.99	32.01	32.62
+ling. & spk.	**27.32**	**29.42**	32.17	32.29

The results on the training set are shown in Table 2, where the ACC and MAP values on both the frame-level (fr.) and the utterance-level (utt.) are reported. It can be seen that with the conditional factors involved, the ACC and MAP values on the training set are significantly improved, and the speaker factor seems to provide more contribution. This improvement on training accuracy demonstrates that with the conditional factors considered, the speech signal can be *explained* much better.

The results on the evaluation set are also shown in Table 2. Again, we observe a clear advantage with the CDF training. Note that involving the two factors does not improve the utterance-level results. This should be attributed to the fact that the DNN models are trained using frame-level data, so may be not fully consistent with the metric of the utterance-level test. Nevertheless, the superiority of the multiple conditional factors can be seen clearly from the frame-level metrics. We note that the discrepancy between the results on the training set and the evaluation set is not due to over-fitting caused by the CDF approach; it simply reflects the mismatch between the training set and evaluation set, hence the difficulty of the task itself. Actually, the results shown here are highly competitive: it beats the MEC 2016 baseline [21] by a large margin, even though we used the 8k Hz data rather than the original 16k Hz data.

6.5. Spectrum reconstruction

In the last experiment, we use the linguistic factor, speaker factor and emotion factor to reconstruct the original speech signal, using the convolutional model shown in Fig. 2. The model is trained using the CHEAVD database. During the training processing, the averaged frame-level reconstruction loss (square error) on the validation set is reduced from 15285.70 to 192.50, and the loss on the evaluation set with the trained model is 196.56. Fig. 3 shows a reconstruction example, where the utterance is selected from the test set of the CHEAVD database. It can be seen that these three factors (linguistic, speaker, emotion) can reconstruct the original spectrum pretty well. Listening tests show that the reconstruction quality is rather good and it is hard for human listeners to tell the difference between the reconstructed speech and the original speech. More examples can be found in the project web site *http://project.cslt.org*. The success of this 'deep reconstruction' indicates that the CDF approach not only factorizes speech signals into task-oriented informative factors, but also preserves most of the information of the speech during the factorization. Moreover, it demonstrates that the cascaded convolution view (Eq. 1) is largely correct. This essentially provides a new vocoder that decomposes speech signals into a sequential convolution of task-oriented factors, which is fundamentally different from the classical source-filter model.

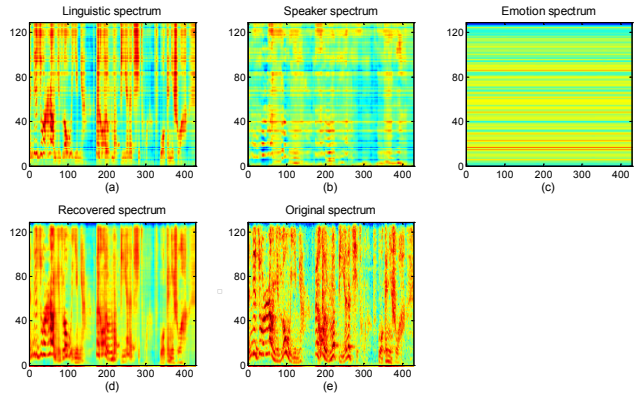

Fig. 3. An example of spectrum reconstruction from the linguistic, speaker and emotion factors.

7. CONCLUSIONS

This paper presented a cascaded deep factorization (CDF) approach to factorize speech signals into individual task-oriented informative factors. Our experiments demonstrated that speech signals can be well factorized at the frame level by the CDF approach, and speech signals can be largely reconstructed using deep neural models from the CDF-derived factors. Moreover, the results on the emotion recognition task demonstrated that the CDF approach is particularly valuable for learning and inferring less significant factors of speech signals.

8. REFERENCES

[1] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang, *Springer handbook of speech processing*, Springer Science & Business Media, 2007.

[2] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] 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.

[4] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[5] Lantian Li, Yixiang Chen, Ying Shi, Zhiyuan Tang, and Dong Wang, "Deep speaker feature learning for text-independent speaker verification," in *INTERSPEECH*, 2017, pp. 1542–1546.

[6] Miao Zhang, Yixiang Chen, Lantian Li, and Dong Wang, "Speaker recognition with cough, laugh and" wei"," *arXiv preprint arXiv:1706.07860*, 2017.

[7] Lantian Li, Dong Wang, Askar Rozi, and Thomas Fang Zheng, "Cross-lingual speaker verification with deep feature learning," *arXiv preprint arXiv:1706.07861*, 2017.

[8] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[9] Dong Yu and Li Deng, *Automatic speech recognition: A deep learning approach*, Springer, 2014.

[10] Zhiyuan Tang, Lantian Li, Dong Wang, and Ravichander Vipperla, "Collaborative joint training with multitask recurrent model for speech and speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 493–504, 2017.

[11] Hiroya Fujisaki, "Communication between minds: The ultimate goal of speech communication and the target of research for the next half-century," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 3025–3025, 1998.

[12] Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi, *Computing prosody: computational models for processing spontaneous speech*, Springer Science & Business Media, 2012.

[13] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1695–1699.

[14] Patrick Kenny, Vishwa Gupta, Themos Stafylakis, P Ouellet, and J Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. Odyssey*, 2014, pp. 293–298.

[15] Rich Caruana, "Multitask learning," in *Learning to learn*, pp. 95–133. Springer, 1998.

[16] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[17] Dong Wang and Thomas Fang Zheng, "Transfer learning for speech and language processing," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015, pp. 1225–1237.

[18] Yanmin Qian, Tian Tan, and Dong Yu, "Neural network based multi-factor aware joint training for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2231–2240, 2016.

[19] Xiangang Li and Xihong Wu, "Modeling speaker variability using long short-term memory networks for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] Wei Bao, Ya Li, Mingliang Gu, Minghao Yang, Hao Li, Linlin Chao, and Jianhua Tao, "Building a chinese natural emotional audio-visual database," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 583–587.

[21] Ya Li, Jianhua Tao, Björn Schuller, Shiguang Shan, Dongmei Jiang, and Jia Jia, "Mec 2016: the multimodal emotion recognition challenge of ccpr 2016," in *Chinese Conference on Pattern Recognition*. Springer, 2016, pp. 667–678.

[22] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[23] Sergey Ioffe, "Probabilistic linear discriminant analysis," *Computer Vision–ECCV 2006*, pp. 531–542, 2006.