IDIAP RESEARCH REPORT

DNN BASED SPEAKER EMBEDDING USING CONTENT INFORMATION FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Subhadeep Dey Takafumi Koshinaka
Petr Motlicek Srikanth Madikeri

Idiap-RR-06-2018

MAY 2018

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

DNN BASED SPEAKER EMBEDDING USING CONTENT INFORMATION FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Subhadeep Dey[1,2], *Takafumi Koshinaka*[3], *Petr Motlicek*[1] *and Srikanth Madikeri*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[3]NEC Corporation
subhadeep.dey@idiap.ch, koshinak@ap.jp.nec.com, petr.motlicek@idiap.ch, srikanth.madikeri@idiap.ch

ABSTRACT

In this paper, we are interested in exploring Deep Neural Network (DNN) based speaker embedding for Random-digit task using content information. To this end, a technique is applied to automatically select common phonetic units between the enrollment and test data to produce speaker verification scores. Furthermore, a novel approach is proposed to incorporate content information in the DNN directly. It is hypothesized that features extracted using this DNN will be helpful for the task. Experiments on the RSR dataset show that the proposed method outperforms the baseline i-vector system by 43% relative equal error rate.

Index Terms— speaker verification, speaker embedding, i-vectors, content mismatch

1. INTRODUCTION

In the last decade, the i-vector approach has shown to be a dominant technique for text-independent Speaker Verification (SV). It assumes that the speaker variabilities lie in a fixed-dimensional subspace [1]. A back-end classifier, such as Probabilistic Linear Discriminant Analysis (PLDA) is applied on the i-vectors for producing SV scores. This approach provides good performance for text-dependent scenario as well [2].

Text-dependent SV is usually implemented using fixed pass-phrases. In this paper, we are interested in Random-digit based text-dependent task which puts less constraints on the speaker. In this scenario, the speaker pronounces a permutation of ten digits during enrollment while the test data consists of five digit string. This scenario is assumed to be robust to replay attack [3].

In literature, various techniques employing i-vector and Joint Factor Analysis (JFA) have been explored for the Random-digit [4, 5]. In [4], parameters of the i-vector model are estimated with the content information of the speech signal. Finally, PLDA model is trained with the content specific vectors. A significant gain in performance is reported as compared to the baseline system. In [6], the JFA model is trained on the segmented digit strings. Finally, the SV scores are obtained by the linear combination of the individual digits.

In another direction, modelling speakers with a speaker discriminative Deep Neural Network (DNN) has shown good performance for SV [7, 8]. Motivated by the success of DNNs in the context of speaker, speech [9, 10] and image recognition tasks, we explore the application of DNNs for the Random-digit task. We believe that the DNN based speaker embedding features can be useful for representing the invariant speaker characteristics.

In [8], a DNN was trained to predict speaker labels for an input speech frame. During evaluation, the activations of the last hidden layer are accumulated over an utterance to obtain speaker representation, referred to as d-Vector. A PLDA is trained as the classifier to provide SV scores. Evaluation on proprietary text-dependent data-set indicates that this technique achieves competitive results as the state-of-the-art SV system. An approach to utterance embedding employing triplet-loss has been successfully applied for SV [11, 12]. This approach is considered as one of the baseline systems in this paper. Triplet-loss involves minimizing the Euclidean distance between same-speaker and maximizing the distance for different-speaker embedding simultaneously. An advantage of this network is that during evaluation, it can be used to obtain speaker similarity directly without training a PLDA separately. However, a limitation of the d-Vector and triplet-loss based approaches is that they ignore the content-information of the speech signal completely. Many studies have shown that exploiting the phonetic variability can significantly enhance the performance of SV systems [5, 4, 13, 14, 15]. Motivated by these results, we incorporated this information on top of DNN based speaker embedding (from the d-Vector and triplet-loss network) for the Random-digit task using content-matching [5]. Content-matching refers to the process of selecting common set of phonetic units between enrollment and test utterance for obtaining speaker similarities. In [5], online i-vectors (i.e. i-vectors estimated for every frame of speech) were used as used as features for performing content-matching. However, a large training corpora (such as Fisher and Switchboard datasets) was required to obtain the reported performance. In this paper, we focus only on the RSR training data for system development. We explore the application of DNN in this scenario to obtain speaker discriminative features. Furthermore, we propose to extract speaker-phonetic features by incorporating content information in the DNN framework of d-Vector and triplet-loss. Experiments are performed on the RSR digit evaluation set. The obtained results significantly outperform the baseline system by 43% relative in Equal Error Rate (EER).

The paper is organized as follows. Section 2 describes the baseline DNN based speaker embedding approaches considered in this paper, while Section 3 describes the content matching and the proposed technique. Sections 4 and 5 present the experimental setup for evaluating the system and discuss the achieved results by various systems. Finally, the paper is concluded in Section 6.

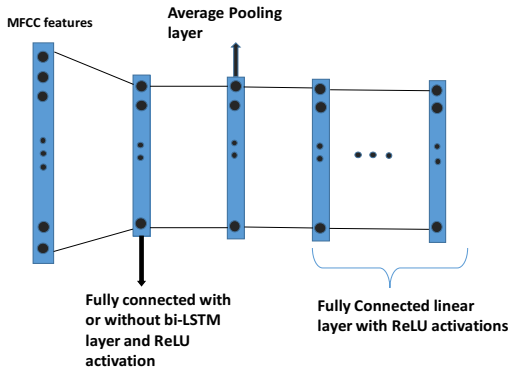

Fig. 1: The neural network architecture of triplet-loss approach for text-dependent SV.

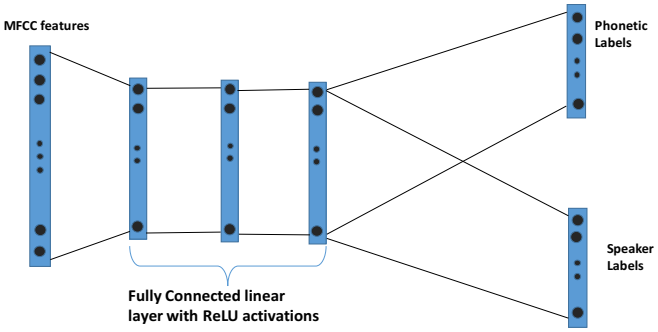

Fig. 2: Multi-task learning for Random-digit task.

2. BASELINE SYSTEMS

Modelling speakers with DNN based speaker-discriminative loss has shown to be beneficial for SV [8, 12]. In this paper, we consider two such successful approaches employing DNN based speaker embedding, namely, (i) d-Vector, and (ii) triplet-loss.

2.1. d-Vector

This approach was first proposed for phrase based text-dependent SV [8]. It trains a DNN that predicts a speaker label with an input speech frame (with context of frames appended to it). The hidden layers of the DNN employ ReLU activation function. The d-Vector representation for an utterance is obtained by averaging the output of the last hidden layer.

2.2. Triplet-loss

Triplet-loss network has been successfully used for speaker diarization and speaker recognition [12, 11, 16, 17]. During training, three utterances, referred to as triplet, $\tau = (\mathbf{X}^a, \mathbf{X}^p, \mathbf{X}^n)$ are fed as input. The utterances are referred to as the anchor, positive and negative examples such that ($\mathbf{X}^a$, $\mathbf{X}^p$) belong to the same speaker while ($\mathbf{X}^a$, $\mathbf{X}^n$) are from different speakers. Thus, the loss function (L) of the network involving the triplet, also known as triplet-loss, aims at minimizing the distance between the embeddings of the anchor and positive, while maximizing the distance between anchor and negative, as given by the following equation:

$$\mathrm{L}(\mathbf{X}^a, \mathbf{X}^p, \mathbf{X}^n) = \|\mathbf{f}(\mathbf{X}^a) - \mathbf{f}(\mathbf{X}^p)\|_2^2 - \|\mathbf{f}(\mathbf{X}^a) - \mathbf{f}(\mathbf{X}^n)\|_2^2, \quad (1)$$

where $\|.\|_2^2$ is the Euclidean norm and $\mathbf{f}$ is the embedding of an utterance produced by the neural network. The network is trained with such triplets τ, so that the triplet-loss (L) is positive. Triplet mining is an important aspect of this approach.

We now describe the architecture of triplet-loss network (as illustrated in Figure 1), which was first proposed in [11] for speaker speaker diarization. We found that the performance of SV system is better on using this architecture compared to a similar network in [7]. The first layer consists of either a bi-directional Long Short Term Memory (bi-LSTM) or fully connected (FC) with ReLU activation function to produce speaker embedding per frame [11, 7]. The bi-LSTM layer takes a single frame whereas the FC layer requires a context of speech frames as input [7]. The second layer called Average Pooling, accumulates the activations from last layer to produce one vector corresponding to an utterance [18, 11]. The subsequent layers consist of FC layers (only one layer is used in our experiments). Finally, length normalization is applied to constrain the output in a fixed dimensional hyper-sphere.

3. CONTENT MATCHING AND SPEAKER-PHONETIC EMBEDDING

The d-Vector and triplet-loss approaches do not use content information for scoring. However, we believe that exploiting this information in addition to speaker embedding would help improve SV. In this regard, we apply content matching for the Random-digit task [5]. Content-matching aims to transform the content of the enrollment data to match the test utterance. Speaker embedding for each frame of an utterance is obtained from the d-Vector and triplet-loss (outputs are collected before the Average Pooling layer of Figure 1) networks for performing content-matching. In [2, 5], i-vector-PLDA features that were trained to discriminate speaker and content were shown to be useful for text-dependent SV. Similarly, we also present an approach to incorporate phonetic information in these networks in order to obtain features that contain both speaker and content information.

3.1. Content matching

It has been shown that exploiting content information can significantly improve the performance of text-dependent SV [5]. In particular, the knowledge of phonetic units co-occurring between enrollment and test data has shown to improve the i-vector system. Content-matching is employed on the speaker embedding per frame (with the d-Vector and triplet-loss network). Similarity scores between enrollment and test data is given by the following equation:

$$s(\mathbf{H_e}, \mathbf{H_t}) = \frac{1}{C} \sum_j min_i d(\mathbf{h}_{e,i}, \mathbf{h}_{t,j}), \quad (2)$$

where $\mathbf{H_e} = \{\mathbf{h}_{e,1}, \mathbf{h}_{e,2}, \cdots, \mathbf{h}_{e,i}, \cdots, \mathbf{h}_{e,R}\}$ and $\mathbf{H_t} = \{\mathbf{h}_{t,1}, \mathbf{h}_{t,2}, \cdots, \mathbf{h}_{t,j}, \cdots, \mathbf{h}_{t,C}\}$ represent set of speaker embeddings per frame for the enrollment and test data, d($\mathbf{h}_{e,i}$, $\mathbf{h}_{t,j}$) computes the distance between the speaker embeddings $\mathbf{h}_{e,i}$ and $\mathbf{h}_{t,j}$. The score $s(\mathbf{H}_e, \mathbf{H}_t)$ represents the accumulated distance between the closest phone units. It is to be noted that this function $s(\mathbf{H}_e, \mathbf{H}_t)$ is different from Dynamic Time Warping in that it does not force locality constraints [19]. A PLDA model is applied to compute the distance (d($\mathbf{h}_{e,i}$, $\mathbf{h}_{t,j}$)) between two speaker embeddings. The details of content matching can be found in [5]. In the next section, approaches to train DNN with phonetic information have been explored.

3.2. Incorporating content information

It has been shown in literature that multi-task learning improves performance for a variety of tasks, like image, speech, speaker recognition [20, 21, 22]. We hypothesize that training the DNN with content information (i.e. extracting speaker-phonetic features from DNN) would be helpful for the Random-digit task. For the d-Vector framework, phonetic information is incorporated by multi-task learning as shown in Figure 2. Multi-task learning involves the joint optimization of the speaker and phonetic loss. After training, the activations from the last hidden layer of Figure 2 are used to represent speaker-phonetic features for performing content matching.

For triplet loss network, phonetic information is applied by using first order statistics of hidden activations instead of Average Pooling layer (Figure 1). These statistics have been used successfully to train a Siamese network [23]. First order statistics summarize the contribution of speakers per phone. The first order statistics ($\mathbf{m}_c$) of an utterance $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M\}$ is computed as follows:

$$\mathbf{m}_c = \sum_i \mathbf{f}(\mathbf{x_i}) \mathbb{1}(\mathbf{x}_i \in c),$$

where $\mathbb{1}(\mathbf{x}_i \in c)$ is an indicator function that outputs one if i^{th} frame is assigned to c^{th} phonetic unit. To obtain the first order statistics, a state-of-the-art automatic speech recognizer is applied to align the development data with mono-phone units. The modified triplet loss function minimizes the embedding of anchor, positive and negative utterances based on the first order statistics, $\mathbf{m}_c$ (similar to the loss function in [23]). The loss function is fully differentiable and the gradients can be estimated efficiently with back propagation algorithm. Once the network has been trained, the outputs after the first layer (bi-LSTM or FC) of Figure 1 are used to perform content matching.

4. EXPERIMENTAL SETUP

In this section, experimental setup of the baseline and the proposed systems are described.

4.1. Evaluation and Training Data

We performed experiments on Part 3 portion of the RSR2015 dataset [3, 24, 25], restricting to female speakers only, which is a Random-digit task. This part comprises 49 speakers uttering random sequence of digits. The enrollment data consists of an user pronouncing ten digits while the test utterance consists of 5 digits. The average duration of the enrollment and test data is 9 s while the test is 3 s respectively. The total number of target trials being 5'283 and 253'584 impostor trials. We used 61K utterances from development and background part of Part 1 to 3 consisting of speech segments spoken by 94 speakers.

4.2. i-vector system

The front-end SV system extracts Mel Frequency Cepstral Coefficients (MFCC) of 20 dimensions from 25 ms frame of speech signal with 10 ms sliding window with the delta and double delta features appended to it. Short time gaussianization is applied to the features using a 3 s sliding window [26]. We trained a 512 mixture Universal Background Model (UBM) on the training data and 200 dimensional i-vector extractor is trained subsequently. Finally, a PLDA is trained as part of the standard recipe of text-independent system with speaker labels of training data [27, 28].

Table 1: Performance of the various baseline systems in terms of EER(%) on RSR2015 Random-digit task. The i-vector system performs better than the DNN based speaker embedding systems. ED refers to Euclidean distance.

Systems	Loss	Architecture	Classifier	EER (%)
i-vector	-	-	PLDA	**11.8**
d-Vector	CE	FC	PLDA	12.3
Multi-task	CE	FC	PLDA	12.7
Triplet	Triplet-loss	bi-LSTM+FC	PLDA	15.2
Triplet	Triplet-loss	bi-LSTM+FC	ED	23.2
Triplet	Triplet-loss	FC+FC	PLDA	17.3
Triplet	Triplet-loss	FC+FC	ED	25.2

4.3. Speaker embedding using the d-Vector and triplet loss network

For the d-Vector, we trained a single layer FC based system with the training data of RSR2015. We used only 940 utterances as the cross-validation data from the 94 speakers. We obtained 100% accuracy on the training and development data using the Cross Entropy (CE) loss function.

For the triplet network, we use offline sampling approach [29]. At any epoch, we generate triplets ($\mathbf{X}^a$, $\mathbf{X}^p$, $\mathbf{X}^n$) such that the phonetic content of these utterances ($\mathbf{X}^a$, $\mathbf{X}^p$, $\mathbf{X}^n$) has maximal overlap. The rationale of choosing this approach is to select difficult negative instances. This leads to creating a total of 200K triplets per epochs. We randomly choose a subset of these triplets to train the triplet loss network. A learning rate of 0.001 was used throughout the experiments. A 1K dimensional hidden layer is used in all the experiments. Pytorch was used for performing the experiments [30]. The performances of various systems are reported in terms of EER.

5. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we describe the results obtained with the baseline and the proposed system. We evaluated the performance of the following systems on the Random digit task:

- **i-vector**: This is the conventional i-vector system using Gaussian Mixture Model (GMM).
- **d-Vector**: A FC hidden layer as used as the network architecture for obtaining d-Vectors. Section 2.1 describes the conventional technique to apply d-Vectors. The baseline d-Vector system employs a PLDA model for scoring. The content matching algorithm uses speaker embedding per frame to produce similarity scores.
- **Triplet**: This approach optimizes the triplet-loss function on three utterances. The conventional approach to using triplet-loss network is described in Section 2.2. This technique uses a bi-LSTM (or a FC) and a FC layer. This network can be applied to compute end-to-end scores (shown as the Euclidean distance in Table 1). Speaker representation of an utterance is obtained by collecting the activations after the Average Pooling layer (See Figure 1). Furthermore, a PLDA model is trained on these representations. The content matching algorithm is applied on the output obtained before Average Pooling layer of Figure 1. The proposed triplet-loss network applying first order statistics (as described in Section 3.2) is referred by **Triplet-Stats**.

Table 2: Performance of the systems using content matching on RSR2015 Random-digit task in terms of EER(%). The baseline i-vector system provides an EER of 11.8%.

Systems	Architecture	EER
d-Vector	FC	9.7
Triplet	FC + FC	9.4
Triplet	bi-LSTM + FC	**6.7**
Neural Networks trained with phonetic information		
Multi-task	FC	7.7
Triplet-Stats	bi-LSTM + FC	13.4

- **Multi-task**: This is the multi-task learning framework involving minimizing the speaker and phonetic loss (using the **d-Vector** approach) as described in Section 3.2. We used only one hidden FC layer with ReLU activation function. For the baseline system employing multi-task learning, hidden activations from last layer are averaged to obtain speaker representation. A PLDA is further trained on these representations. For content matching, hidden activations per frame are applied for obtaining SV scores.

5.1. Baseline SV systems

Table 1 shows the performance of i-vector and DNN speaker embedding based SV systems. The performance of the i-vector system is comparable to the results published in literature [5]. From Table 1, we observe that the simple d-Vector approach performs better than the triplet loss network. Furthermore, we observe that multi-task training provides worse performance than the d-Vector. Thus, effects of averaging of hidden activation are worse in multi-task than d-Vectors. Performance of the d-Vector approach is worse than the baseline i-vector PLDA system.

For the triplet based network, we observe that the performance is worse compared to that of the baseline i-vector and the d-Vector approaches. It is to be noted that Triplet system provides an EER of 23.2% using Euclidean distance or end-to-end loss only (PLDA was not applied in this system). An explanation of the poor performance of the Triplet approach is that it requires large speaker population to provide results comparable to i-vector system [7, 23].

5.2. Proposed approach

Table 2 shows the results of the content matching using speaker embeddings from the different networks. The results indicate that performance of the systems (d-Vector and Triplet) dramatically increase, when content matching is applied. It highlights the importance of using common phones for obtaining speaker similarities. Furthermore, we observe that the triplet loss network employing bi-LSTM and FC layers performs significantly better the other approaches using content matching. It outperforms the baseline i-vector system by 43% relative EER (11.8% to 6.7% absolute).

Integrating phonetic information in the DNN has shown to provide promising results. In particular, the multi-task training in the d-Vector significantly outperforms the speaker-loss by 20% relative EER (9.7% to 7.7% absolute). This system outperforms the baseline i-vector system by 34% relative EER (11.8% to 7.7% absolute). However, performance of triplet-loss network using content information degrades significantly . A reason of this could be that the phonetic information is fully deterministic and can not be optimized by the neural network.

6. CONCLUSIONS

In this paper, various approaches to apply DNN based speaker embedding features for Random-digit task were explored using content information. In this regard, we considered two DNN approaches to speaker verification, namely (i) d-Vector, and (ii) triplet-loss. We applied content matching to produce speaker similarities by selecting common phonetic units between the enrollment and the test data. We also proposed an approach to incorporate content information in the DNN by multi-task learning and computing the first order statistics. We observed that the estimating speaker embedding from the triplet-network performs the best and outperforms the baseline i-vector system by 43% relative EER. Furthermore, speaker embedding features obtained from multi-task trained network performed better than the d-Vector.

7. ACKNOWLEDGEMENTS

This work was supported by the EU FP7 project Speaker Identification Integrated Project (SIIP) and NEC Corporation, Japan.

8. REFERENCES

[1] Daniel Garcia Romero and Carol Y. Espy Wilson, "Analysis of ivector length normalization in speaker recognition systems," in *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27 to 31, 2011*, 2011, pp. 249–252.

[2] S. Dey, S. Madikeri, M. Ferras, and P. Motlicek, "Deep neural network based posteriors for text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, March 2016.

[3] Anthony Larcher, Kong-Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[4] Liping Chen, Kong Aik Lee, Bin Ma, Wu Guo, Haizhou Li, and Li-Rong Dai, "Phone-centric local variability vector for text-constrained speaker verification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] Subhadeep Dey, Srikanth Madikeri, Petr Motlicek, and Marc Ferras, "Content normalization for text-dependent speaker verification," *Proc. Interspeech 2017*, pp. 1482–1486, 2017.

[6] Themos Stafylakis, Md Jahangir Alam, and Patrick Kenny, "Text-dependent speaker recognition with random digit strings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 7, pp. 1194–1203, 2016.

[7] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5115–5119.

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

[9] David Imseng, Petr Motlicek, Philip N. Garner, and Hervé Bourlard, "Impact of deep mlp architecture on different acoustic modeling techniques for under-resourced speech recognition," in *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, Dec. 2013.

[10] Herve Bourlard, John Dines, Mathew Magiami-Doss, Philip Garner, David Imseng, Petr Motlicek, Hui Laing, Lakshmi Saheer, and Fabio Valente, "Current trends in multilingual speech processing," *Sadhana*, vol. 36, no. 5, pp. 885–915, Oct 2011.

[11] Hervé Bredin, "Tristounet: triplet loss for speaker turn embedding," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5430–5434.

[12] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[13] Subhadeep Dey, Petr Motlicek, Srikanth Madikeri, and Marc Ferras, "Exploiting sequence information for text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. Ieee, 2017, pp. 5370–5374.

[14] Petr Motlicek, Subhadeep Dey, Srikanth Madikeri, and Lukas Burget, "Employment of subspace gaussian mixture models in speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4445–4449.

[15] Subhadeep Dey, Petr Motlicek, Srikanth Madikeri, and Marc Ferras, "Template-matching for text-dependent speaker verification," *Speech Communication*, vol. 88, pp. 96–105, 2017.

[16] Fabio Valente, Deepu Vijayasenan, and Petr Motlicek, "Speaker diarization of meetings based on speaker role n-gram models," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[17] Fabio Valente, Petr Motlicek, and Deepu Vijayasenan, "Variational bayesian speaker diarization of meeting recordings," in *Proceedings of ICASSP*, 0 2010.

[18] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," .

[19] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.

[20] Ronan Collobert and Jason Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 160–167.

[21] Yoshua Bengio et al., "Learning deep architectures for ai," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[22] Alexandros Lazaridis, Ivan Himawan, Petr Motlicek, Iosif Mporas, and Philip N Garner, "Investigating cross-lingual multi-level adaptive networks: The importance of the correlation of source and target languages," in *Proceedings of the International Workshop on Spoken Language Translation*, 2016, number EPFL-CONF-223756.

[23] Shi-Xiong Zhang, Zhuo Chen, Yong Zhao, Jinyu Li, and Yifan Gong, "End-to-end attention based text-dependent speaker verification," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 171–178.

[24] A. Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Modelling the alternative hypothesis for text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 734–738.

[25] A. Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7673–7677.

[26] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," 2001, pp. 213–218, In Proc. of Speaker Odyssey.

[27] Srikanth Madikeri, Subhadeep Dey, Petr Motlicek, and Marc Ferras, "Implementation of the standard i-vector system for the kaldi speech recognition toolkit," Tech. Rep., Idiap, 2016.

[28] Subhadeep Dey, Srikanth Madikeri, and Petr Motlicek, "Information theoretic clustering for unsupervised domain-adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5580–5584.

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[30] Pytorch, "https://github.com/pytorch," 2017.