Analysis of Complementary Information Sources in the Speaker Embeddings Framework

Mahesh Kumar Nandwana, Mitchell McLaren, Diego Castan, Julien van Hout, Aaron Lawson

Speech Technology and Research Laboratory, SRI International, Menlo Park, California, USA

{mahesh.nandwana, mitchell.mclaren, diego.castan, julien.vanhout, aaron.lawson}@sri.com

Abstract

Deep neural network (DNN)-based speaker embeddings have resulted in new, state-of-the-art text-independent speaker recognition technology. However, very limited effort has been made to understand DNN speaker embeddings. In this study, our aim is analyzing the behavior of the speaker recognition systems based on speaker embeddings toward different front-end features, including the standard Mel frequency cepstral coefficients (MFCC), as well as power normalized cepstral coefficients (PNCC), and perceptual linear prediction (PLP). Using a speaker recognition system based on DNN speaker embeddings and probabilistic linear discriminant analysis (PLDA), we compared different approaches to leveraging complementary information using score-, embeddings-, and feature-level combination. We report our results for Speakers in the Wild (SITW) and NIST SRE 2016 datasets. We found that first and second embeddings layers are complementary in nature. By applying score and embedding-level fusion we demonstrate relative improvements in equal error rate of 17% on NIST SRE 2016 and 10% on SITW over the baseline system.

Index Terms: Speaker recognition, deep neural networks, speaker embeddings, score fusion, feature fusion, logistic regression.

1. Introduction

Deep neural networks (DNN) have recently revolutionized the area of speech technology. Their introduction in automatic speech recognition (ASR), speaker and language recognition [1, 2], speech activity detection (SAD), emotion recognition, and speech enhancement systems has resulted in significant performance gains over prior technologies.

In the area of speaker recognition, recent advances have seen DNNs replacing the universal background model (UBM) for frame alignment [3] and the use of DNNs for bottleneck feature extraction [4]. More recently, DNNs have been directly used to capture speaker characteristics by training to discriminate between speakers, which results in a fixed-dimensional speaker representation extracted from one of its hidden layers [5, 6, 7, 8]. This representation is referred to as a speaker embedding. Initial research in this area was primarily concentrated on *text-dependent* speaker recognition, in which d-vectors were used for text-dependent speaker recognition [6]. Of great interest to the speaker recognition community was the generalization of this technology to the *text-independent* case [7, 8]. This context is the focus of this article, with speaker embeddings used to replace the i-vectors employed in traditional speaker recognition systems [9] to offer a considerable improvement to system robustness [8]. Besides the initial studies on speaker embeddings, the topic of data augmentation to improve system performance has also been a main focus [1, 10, 11], as well as its application to speaker diarization [12].

In this paper, our goal is to gain a deeper understanding of speaker embeddings with respect to leveraging complementary information to improve system performance. This topic is highly relevant to real-world users of speaker embeddings, in which computational needs must be balanced against performance, and in which researchers are interested in the combination of multiple systems to maximize accuracy in speaker recognition evaluations. Initial work in the area of embedding-system combination was presented in [8], in which scores from two different embeddings layers from the same DNN were averaged. We extend this research by comparing the robustness of different front-end features for the speaker embedding extractor. We also explore different output layers of the embedding extractor network and analyze the strengths and weaknesses of each. This is followed by fusion of different information sources at the score, embeddings, and feature-levels to understand where complementarity can be best exploited. Our experiments are conducted using NIST SRE 2016 [13] and the publicly available dataset Speakers in the Wild (SITW) [14].

The remainder of this paper is organized as follows: First, the speaker recognition system used in this study is detailed. This is followed by a brief overview of approaches to leverage complementary information in speaker recognition systems in Section 3. Next, Sections 4 and 5 describe the evaluation protocol and results, respectively. Section 6 presents the conclusion of the study and provides directions for future research.

2. Speaker Recognition System

In this section, we detail our speaker recognition system that is used throughout this study. The main components of our system include speech activity detection (SAD), a DNN-based embedding extractor, and a probabilistic linear discriminant analysis (PLDA) classifier. While important for real system use, the aspect of system calibration is not covered in the scope of this article and will be left for future research.

2.1. Speech Activity Detection

In our previous work [10], we investigated the impact of speech activity detection (SAD) on the performance of speaker-embeddings-based speaker recognition systems. It was shown that a low SAD threshold during training tended to benefit the embeddings extractor, while maintaining a strict threshold during evaluation was necessary. Our SAD is DNN-based with two hidden layers containing 500 and 100 nodes, respectively. The SAD DNN is trained using 20-dimensional Mel-frequency cepstral coefficients (MFCC) features, stacked with 31 frames. Before training the SAD DNN, the features were mean and variance normalized over a 201-frame window. The threshold for selecting the speech versus non-speech frames was 0.5 for evaluation and -1.5 for DNN training.

2.2. Speaker Embedding Extractor

The architecture of our speaker embeddings extractor DNN follows the Kaldi recipe [15]. This feed-forward DNN is trained to discriminate between speakers. Through the use of a statistics pooling layer, the DNN maps a variable length utterance to a fixed-dimensional embedding.

The embeddings network has five frame-level hidden layers with rectified linear unit (ReLU) activation and batch normalization. The first three layers incrementally add time context with stacking of [-2, -1, 0, 1, 2], [-2, 0, 2], and [-3, 0, 3] instances of the input feature frame. Means and standard deviations of the frame-per-audio segments are stacked using a statistic pooling layer. The final two hidden layers of 512 nodes operate at the segment-level prior to the log-soft-max output layer. A ReLU activation function and batch normalization prior to a layer's output is applied to all layers except the output layer. Speaker embeddings can be extracted either from first or second segment-level hidden layer, each being 512 nodes.

2.3. Training Data

Training data plays an important role in the design of a robust speaker embeddings extractor [1, 10]. For training the embeddings extractor, we follow the recipe in [10] for the raw+CNlowRM system. Specifically, we start with the non-degraded subset of the PRISM training lists [16], which contains 52,456 audio files. We then augment this data with four copies of degraded data using a random selection of audio compression; a random selection of noises at a 5 dB signal-to-noise ratio (SNR); a random selection of instrumental music at a 5 dB SNR; and a random selection of reverberated signals with low reverberation. The raw training data along with the augmented data copies are used to train the DNN for embedding extraction. This results in a total of 262,280 segments from 3,296 speakers. For interested readers, additional details on these degradations and how they were applied can be found in [10].

2.4. Probabilistic Linear Discriminant Analysis (PLDA) Classifier

A PLDA model [17] learns the within-class and across-class variabilities of a large, labeled training set using expectation maximization (EM). We use gender-independent PLDA for all our experiments described herein. Before training the PLDA classifier, the dimensions of the embeddings are reduced to 200 using linear discriminant analysis (LDA), followed by length normalization and mean centering [18] to Gaussianize the distribution of the speaker embeddings. Finally, these normalized speaker embeddings are used by the PLDA classifier to compute a similarity score between speaker embeddings. The full PRISM training lists (including original degradations) [16] with additional transcoded data is used for training the PLDA model.

3. Leveraging Complementary Information in Speaker Recognition

In speaker recognition systems, complementary information is often used to improve the robustness of the system and its ability to generalize to unseen data. In the NIST speaker recognition evaluations (SRE), teams often develop multiple systems, which are complementary in nature and are used in combination to form a robust system [19, 20].

These complementary information sources can be combined in numerous ways. Typically, this fusion is performed at the score-level, which enables each system to be developed independently, at the cost of increased computational needs for the full system. In this work, we investigate different points of fusion within the embeddings architecture to determine how to maximize performance while minimizing the system's computational requirements. To this end, we fuse information from multiple sources by using (i) score-level fusion, (ii) embedding-level fusion, and (iii) feature-level fusion.

3.1. Score-Level Fusion

Score-level fusion is the most widely used technique to leverage complementary information from multiple sub-systems in speaker recognition. In this work, we use two different types of score-level fusion approaches.

3.1.1. Score Averaging

This approach is usually applied where there is an absence of held out data. This type of fusion refers to the equally weighted sum of PLDA scores of multiple sub-systems. This approach was used in [8] for fusion of PLDA scores from i-vector and speaker embedding sub-systems.

3.1.2. Logistic Regression

Logistic regression based score-level fusion typically refers to the application of logistic regression to learn a set of fusion and calibration parameters that optimize the combination of information sources on a held-out dataset [21, 19]. The advantage of logistic regression training is two-fold: First, it improves the discriminative ability of the system. Secondly, it calibrates the output score, so that it functions as a well-calibrated log-likelihood ratio. Additional mathematical details on logistic regression fusion for speaker recognition can be found in [19].

3.2. Embeddings-Level Fusion

Embeddings-level fusion is inspired by our i-vector fusion paradigm [20]. This process involves applying linear discriminant analysis (LDA) to each individual source of information (embeddings) to reduce the embeddings to 200 dimensions, concatenating the reduced embeddings, and then applying a final LDA transform to reduce the concatenated embeddings to 200 dimensions. The embeddings can either be from different output layers of the same embedding-extractor network or from different embeddings extractors.

3.3. Feature-Level Fusion

Different types of acoustic features represent speaker-specific characteristics in a complementary manner [22]. Feature-level fusion is the simple process of concatenating feature vectors from multiple sources. This happens at the early stage as opposed to score-level fusion which is a late fusion. For instance, concatenating two features of 20 dimensions will result in a 40-dimensional fused feature. The fused feature is then fed as an input to the speaker embedding extractor DNN.

4. Evaluation Protocol

We evaluate our speaker recognition system on two different datasets: (i) NIST SRE 2016 and (ii) Speakers in the Wild (SITW). These datasets contain a range of conditions such as channel, language, codecs, noises, reverberation etc. usually observed in the real-world. The performance of the system

Table 1: *Details of different evaluation conditions from NIST SRE 2016 and SITW and number of target/impostor (tgt/imp) trials used in this work.*

Conditions	# Spk	# Tgt	# Imp
sre16-tgl	101	17764	1003568
sre16-yue	100	19298	946098
sre16-all	201	37062	1949666
SITW	180	3654	716933

across these datasets is measured in terms of equal error rate (EER).

NIST SRE 2016: NIST Speaker Recognition Evaluation (SRE) 2016 dataset includes variabilities due to domain/channel and language mismatches. It also has variability in test segment duration which is uniformly distributed between 10s and 60s. The non-English conversational telephone speech (CTS) is recorded over a wide variety of handset types. We report our results on Tagalog (sre16-tgl) and Cantonese (sre16-yue), also referred to as the *major* languages in the evaluation as well as on overall set.

SITW: The Speakers in the Wild (SITW) dataset contains speech samples in English from open-source media [23]. SITW contains naturally occurring noises; reverberation; codec; and channel variability. The enroll and test utterances for SITW vary in duration from 6–240 seconds. We report our results on evaluation set of SITW data.

The details of different evaluation conditions used in the experiments from NIST SRE 2016 and SITW are summarized in Table 1. For training of the calibration parameters in score-level fusion using logistic regression, the development set from NIST SRE 2016 and SITW were used.

5. Results and Analysis

This section investigates how to best exploit the complementary information between embedding systems trained using different acoustic features. We first benchmark three different front-end features. We then explore single system complementarity and multiple feature combination using approaches proposed in Section 3.

5.1. Front-End Features

Prior to investigating how to maximize the complementary information from different acoustic features in the speaker embeddings framework, we provide a benchmark of the features considered in this study. From these results, we can quantify the relative contribution of additional information in the system.

Table 2: *Equal Error Rate (EER) in % of individual acoustic features used in this study. Both first and second embeddings layers were evaluated.*

Condition	1st Embedding			2nd Embedding		
	MFCC	PNCC	PLP	MFCC	PNCC	PLP
sre16-tgl	23.05	**21.03**	22.98	21.89	21.05	24.18
sre16-yue	10.11	**8.25**	11.28	10.01	9.33	12.98
sre16-all	17.88	**16.16**	18.16	17.21	16.2	20.2
SITW	7.22	8.46	8.56	**7.09**	9.14	9.09

We selected three widely used front-end features for our analysis: (i) MFCC, (ii) power normalized cepstral coefficients (PNCC) [24], and (iii) perceptual linear prediction (PLP) [25]. We use 20-dimensional MFCC and PNCC features with the frame rate of 25 ms and the step size of 10 ms. PLP features are 13-dimensional vectors with the same frame rate and step size as MFCC and PNCC. Our baseline system is using MFCC features with embeddings extracted from first output layer, which is similar to [1].

Both embedding layers were benchmarked independently for each of these features, with results summarized in Table 2. Several trends can be observed from these results. First, PNCC provides better performance across SRE 16 conditions using either first or second embedding layer over MFCC and PLP. Second, MFCC features consistently provide better performance using the second embedding layer across all evaluation conditions. From these results, it is clear that no single feature is best suited to either dataset. It is in such scenarios that system fusion aims to harness the most useful information from individual features and provide and better system overall.

Quantifying the difference between features, PNCC features give a relative improvement of 9.6% for SRE 16, while MFCC with second layer embedding resulted in an improvement of more than 6% for SITW set. PLP, on the other hand, offered the worst performance across both datasets

5.2. Single System Complementarity

Due to the benefits of each feature or embedding layer choice being dataset-dependent, we aim to analyze the methods of leveraging complementary information in the embeddings framework as described in Section 3.

Table 3: *Equal Error Rate (EER) in % when combining first and second embeddings layers through score-level or embedding-level fusion.*

Condition	MFCC	PNCC	PLP
Score Averaging			
sre16-tgl	21.83	20.38	22.97
sre16-yue	9.08	**7.85**	11.43
sre16-all	16.83	15.44	18.78
SITW	6.54	8.16	8.13
Logistic Regression			
sre16-tgl	21.86	20.40	22.73
sre16-yue	9.08	7.88	11.07
sre16-all	16.83	15.45	18.19
SITW	**6.49**	8.16	8.10
Embedding-Level combination			
sre16-tgl	21.69	**20.01**	22.09
sre16-yue	9.21	**7.85**	11.43
sre16-all	16.50	**14.83**	17.91
SITW	6.63	8.40	8.70

We first look at how to best exploit complementary information within a single embeddings network. This is done using either score-level fusion or embedding-level fusion of the first and second embeddings layers from a given embeddings network. For score-level fusion we explore both simple averaging as in [8] and logistic regression fusion. Results are reported in Table 3.

Fusion of first and second layer with all three approaches

resulted in the improved performance for all three features over their first layer embedding alone. For SRE 16, embedding-level combination was better than at score-level, whereas for SITW, logistic regression fusion resulted in the best performance. For the sre16-all condition, embedding-level fusion of PNCC features resulted in a 17% relative improvement over MFCC features. For SITW, logistic regression based score fusion improved the performance by 10%. This demonstrates that the first and second output layer embedding are complementary in nature.

For score-level fusion, we observe that both score averaging and logistic regression fusion yield similar results. This shows that for the MFCC and PNCC features, the embeddings layers offer similar levels of useful information in the fusion, while in the case of PLP in which logistic regression was consistently better, the fusion process was able to determine more appropriate weights for each embedding layer than a simple average.

5.3. Multiple Feature Combination

In this section, we expand the combination of systems to include more than one acoustic feature; a process in which more complementary information is expected due to the unique processing of the audio from each feature. To observe the complementarity of acoustic features, we restrict the analysis to the first embeddings layer. Expanding on the combination methods of the previous section, we also analyze feature-level fusion that simply concatenates features at the frame level prior to input to the embeddings DNN.

Table 4: *EER(%) using multiple acoustic features in up to three methods of combination in the embeddings framework.*

Condition	MFCC+PNCC	MFCC+PLP	PNCC+PLP
Score Averaging			
sre16-tgl	21.43	22.48	21.27
sre16-yue	8.39	9.92	8.79
sre16-all	16.43	17.51	16.51
SITW	7.11	6.96	7.77
Logistic Regression			
sre16-tgl	21.09	22.44	20.99
sre16-yue	8.19	10.04	8.28
sre16-all	16.15	17.52	16.12
SITW	6.90	**6.79**	8.46
Embedding-level combination			
sre16-tgl	21.10	22.72	**20.74**
sre16-yue	**8.13**	9.73	8.67
sre16-all	**15.86**	17.77	15.98
SITW	7.14	7.06	8.10
Feature-level combination			
sre16-tgl	22.90	24.04	21.32
sre16-yue	10.21	10.37	9.48
sre16-all	17.93	18.67	16.37
SITW	9.66	7.53	11.36

Table 4 provides results for numerous combination options across embeddings systems. Combination of multiple acoustic feature subsystems resulted in improved performance at score and embedding-level fusion for MFCC and PNCC. However multiple subsystem fusion did not outperform the single system complementarity from both embeddings layers of a single DNN shown in the previous section.

One of the things to note in Table 4 is that feature-level combination did not yield significant gains and in some cases, it resulted in a performance loss over the baseline. This was probably due to that fact that features are highly correlated and may require some type of dimensionality reduction approaches such as principal component analysis (PCA) or LDA to reduce the dimensions before feeding the features to the embedding DNN.

6. Conclusions and Future Work

In this work, we investigated front-end features and fusion approaches for a text-independent speaker recognition system based on deep neural network speaker embeddings. We experimented with three different features and explored fusion at the score-, embeddings-, and feature-level. We found that MFCC and PNCC features are more robust compared to PLP features. The first and second layer embeddings provide information that is complementary. The fact that the second embeddings layer can be extracted from the first with very limited overhead provided motivation for its use. In the context of multiple acoustic features, score- and embedding-level fusion helps increase the overall robustness of speaker recognition systems while feature fusion was found to degrade the performance of the system in some cases. In the future, we plan to explore the feature-level contextualization alongside the internal contextualization of the DNN and leverage side or meta information in the training of the embeddings DNN to better exploit complementary information in a condition-dependent manner. We also plan to explore various dimensionality reduction approaches in order to extract information from feature-level fusion.

7. References

[1] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[2] M. McLaren, M. K. Nandwana, D. Castan, and L. Ferrer, "Approaches to multi-domain language recognition," *Speaker Odyssey*, 2018.

[3] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1695–1699, 2014.

[4] M. McLaren, Y. Lei, and L. Ferrer, "Advances in deep neural network approaches to speaker recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4814–4818, 2015.

[5] K. Chen and A. Salman, "Learning speaker-specific characteristics with a deep neural architecture," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1744–1756, 2011.

[6] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4052–4056, 2014.

[7] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," *IEEE Spoken Language Technology Workshop (SLT)*, pp. 165–170, 2016.

[8] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," *INTERSPEECH-2017*, pp. 999–1003, 2017.

[9] 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.

[10] M. McLaren, D. Castan, M. K. Nandwana, L. Ferrer, and E. Yilmaz, "How to train your speaker embeddings extractor," *Speaker Odyssey*, 2018.

[11] M. K. Nandwana, J. van Hout, M. McLaren, A. Stauffer, C. Richey, A. Lawson, and M. Graciarena, "Robust speaker recognition from distant speech under real reverberant environments using speaker embeddings," *INTERSPEECH-2018*, 2018.

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

[13] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. Greenberg, E. S. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," *INTERSPEECH-2017*, pp. 1353–1357, 2017.

[14] M. McLaren, D. Castan, and L. Ferrer, "Analyzing the effect of channel mismatch on the SRI language recognition evaluation 2015 system," *Speaker Odyssey 2016*, pp. 188–195, 2016.

[15] *NIST SRE 2016 Xvector Recipe*, 2017, https://david-ryan-snyder.github.io/2017/10/04/model_sre16_v2.html.

[16] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot *et al.*, "Promoting robustness for speaker modeling in the community: the PRISM evaluation set," *Proceedings of NIST 2011 workshop*, 2011.

[17] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

[18] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," *INTERSPEECH 2011*, pp. 249–252, 2011.

[19] N. Brummer, J. Cernocky, M. Karafiát, D. A. van Leeuwen, P. Matejka, P. Schwarz, A. Strasheim *et al.*, "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[20] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation," *INTERSPEECH 2013*, pp. 1981–1985, 2011.

[21] D. A. Van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker classification I*, 2007, pp. 330–353.

[22] M. J. Alam, P. Kenny, and T. Stafylakis, "Combining amplitude and phase-based features for speaker verification with short duration utterances," *INTERSPEECH-2015*, pp. 249–253, 2015.

[23] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," *INTERSPEECH-2016*, pp. 818–822, 2016.

[24] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4101–4104, 2012.

[25] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.