MULTI-FEATURE INTEGRATION FOR SPEAKER EMBEDDING EXTRACTION

Sreekanth Sankala , Shaik Mohammad Rafi B, Sri Rama Murty K

Speech Information Processing Lab
Indian Institute of Technology Hyderabad
{ee20resch11011,ee17resch01003,ksrm}@iith.ac.in

ABSTRACT

The performance of the automatic speaker recognition system is becoming more and more accurate, with the advancement in deep learning methods. However, current speaker recognition system performances are subjective to the training conditions, thereby decreasing the performance drastically even on slightly varied test data. A lot of methods such as using various data augmentation structures, various loss functions, and integrating multiple features systems have been proposed and shown a performance improvement. This work focuses on integrating multiple features to improve speaker verification performance. Speaker information is commonly represented in the different kinds of features, where the redundant and irrelevant information such as noise and channel information will affect the dimensions of different features in a different manner. In this work, we intend to maximize the speaker information by reconstructing the extracted speaker information in one feature from the other features while at the same time minimizing the irrelevant information. The experiments with the multi-feature integration model demonstrate improved performance than the stand-alone models by significant margins. Also, the extracted speaker embeddings are found to be noise-robust.

Index Terms— Speaker verification, Feature fusion, Self supervised learning, Deep speaker embeddings.

1. INTRODUCTION

Automatic Speaker Verification (ASV) is the process of validating the identity of a claimed speaker in a given speech segment. Processing a speech segment can provide speaker-specific information such as formants, pitch, behavioral aspects like speaking rate, etc., which are essential for the ASV systems [1]. Predominantly, the ASV systems receive input as the handcrafted features extracted from the speech signal using traditional short-time signal processing techniques. In a low dimensional embedding space, these short-time features are statistically modeled to enhance speaker-specific information, alleviating channel and other environmental effects.[2, 3]. On the other hand, with the advent of deep learning methods, there is rapid improvement in the performance of ASV systems where these handcrafted features are projected into higher-order nonlinear speaker discriminative spaces and scored using a backend technique. The speaker information is distributed throughout the expanse of the spoken utterance and hence modeled as a fixed-length embedding, extracted by temporally aggregating the frame-level handcrafted features [4]. Nonetheless, the direct usage of speech samples as input to the deep network is trending and found to be outperforming the traditional feature-based ASV system [5, 6]. Unlike handcrafted features, raw samples based speech representations are not restricted to the magnitude or phase component of the signal. Typically, they learn the information that relies on the network's final objective. However, there are self-supervised methods that extract speech representations rich in textual as well as speaker information [7]. These representations are later used for the task of speaker recognition and found to be performing well [8].

In spite of such advances in speech feature representations, there is still a gap in the literature to capture complete speaker-specific information from a given utterance. In order to capture the speaker traits, the researchers use a variety of features and speech samples as input to the ASV systems. However, only one kind of such feature representation may not be adequate to capture the complete speaker information, especially using the networks trained in a discriminatory manner. As different features have different speaker-specific properties, noise robustness etc, it is always effective to integrate multiple features for speaker identification. There are numerous works that fuse different kinds of these features to capture speaker-specific evidence from multiple representations of the same speech segment to improve speaker verification performance [9, 10, 11, 12]. Most of these works either combine features at the input or fuse the final scores. However, all these works lack generalization of speaker information across the features while modeling.

In this study, we propose a deep neural network architecture that integrates evidence from multiple representations of the speech segment at frame level and extracts a segment level embedding which adapts the characteristics of the given multiple low-level information sources. Such a multi-feature integration model facilitates the information sharing among different features at the segment level and hence maximizes

the speaker information. Experimental results show that the proposed speaker embeddings are more discriminative and robust to noise conditions. The remainder of the paper is organized as follows, Section 2 introduces the speech features and proposed architecture of multi-feature integration. Sections 3 and 4 discuss the experiments and results respectively. Finally Section 5 concludes the study with future directions.

2. MULTI-FEATURE INTEGRATION EXPERIMENTS

This section will describe the feature extraction and proposed multi-feature pooling method where we will use independent branches for multi-feature processing and a common segment level layer to train the network.

2.1. Speaker information extraction from stand alone features

Typically magnitude-based speech features are prominent in speech applications. Such magnitude information can be extracted by discarding phase using Fourier and analytic signal analysis in frequency and time domains. In the frequency domain, Mel frequency cepstral coefficients (MFCC) are conventionally used features where auditory-like processing is performed on the magnitude (power) spectrum of short-time Fourier transform (STFT) and further compressed with DCT.

In general, speech signals are characterized by a variation of frequency content with time, and the STFT does not show this time-frequency resolution variations effectively. Nevertheless, the instantaneous amplitude and phase information can be estimated by the AM-FM decomposition of the signal in the analytic domain. Specifically, the instantaneous amplitude referred to as the Hilbert envelope can be approximated using a linear prediction on the DCT components representing critical sub-bands of a signal [13]. The features extracted by processing such Hilbert envelope are called *frequency domain linear prediction* (FDLP) features. Hence we chose MFCC and FDLP features extracted from Fourier and analytic domain magnitude as standard hand-crafted features. Both features are found to be potential in improving the performance of speaker recognition systems [13].

On the other hand, the self-supervised speech representations extracted from raw speech samples are prevalent in most of speech applications. The self-supervised models transform speech signals into latent speech representations over a given context and are trained to solve a reconstruction loss or contrastive loss. In recent times, the pre-trained wav2vec model is widely used as a front-end feature extractor for various speech applications.

2.2. Multi feature pooling and speaker embedding extraction

The magnitude-based information from the MFCC and FDLP and the self-supervised speech representations from the raw speech samples are individually processed at the frame level. We extract a segment-level speaker embedding that is aware of speaker traits present in all three kinds of features. Motivated by the expeditious success of the x-vector architecture, we follow the similar paradigm of processing frame-level information and extracting an utterance-level representation by aggregating the frame-level representations.

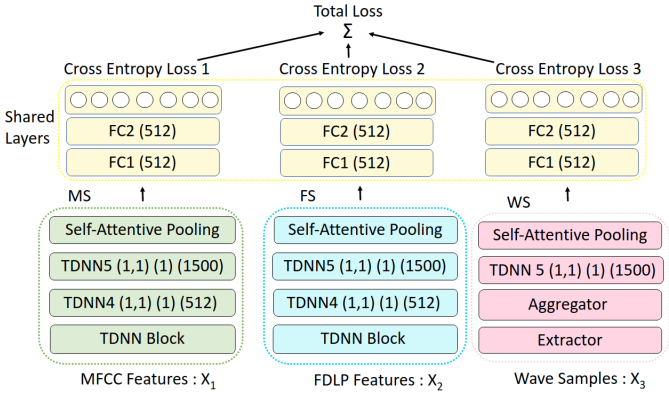


Fig. 1: Proposed speaker embedding extraction model with multi feature fusion.

The proposed deep learning framework is shown in Fig. 1. In order to collect speaker-rich information from multiple feature representations, independent frame-level encoders are dedicated to process each kind of feature input. Out of the three, two streams are similar with Time-delay neural network (TDNN) layers that perform dilated convolutions to capture temporal context information of a given segment for processing MFCC and FDLP features. And the pre-trained wav2vec model as given in [7] is employed as the third feature processing stream. In order to match the dimensionality with the other two feature encoders, an additional dense layer is appended on top of the wav2vec model. Every feature encoder has a corresponding temporal pooling layer that computes statistics (mean and variance vectors with respect to time) to get the gross segment level representation. As different channels carry distinct information across the feature spaces, a channel-wise attention mechanism [14] is integrated at frame level to perceive and emphasize speaker information across the channels. All three independent feature encoders share the segment-level layers of the network. The cross-entropy loss is computed for each kind of feature, and the three losses are summed to get the final loss of the network.

$$\mathcal{L}_{final} = \mathcal{L}_m\{\theta_m, \theta_s\} + \mathcal{L}_f\{\theta_f, \theta_s\} + \mathcal{L}_w\{\theta_w, \theta_s\} \quad (1)$$

$\theta_m, \theta_f, \theta_w$ are frame level feature encoder parameters of MFCC, FDLP and Wav2vec respectively and θ_s denote the segment level network parameters. The training of such a network processes the collective information distributed across the features at the frame level into a segment level speaker embedding. While inference, the proposed model is flexible to exploit speaker information from any of the three features because of the shared segment level layer. The shared segment level layer can be mounted on any three feature encoders, and the individual performance is evaluated. In the next section, we show the improved performance of ASV systems over the stand-alone systems with the help of our experiments.

3. EXPERIMENTS

3.1. Dataset

In this work, VoxCeleb1 development data [15] along with the 5-fold data augmentation [16] is used to train all the speaker embedding models. 40 speakers in the VoxCeleb1 test dataset are used to evaluate the models. Along with the VoxCeleb1 test for evaluation, Voxceleb-1 noisy data, synthesized using Microsoft scalable noise dataset [17] is used to evaluate the robustness of the system in noise conditions.

3.2. Architecture and Training Details

Each of the MFCC and FDLP frame-level networks consists of five TDNNs layers similar to that of the x-vector model [4]. The wav2vec model consists of two blocks of deep neural networks, the feature extractor, and the feature aggregator. Feature extractor comprises five convolutional layers with kernel sizes (10,8,4,4,4) and strides (5,4,2,2,2). Feature aggregator network has nine convolutional layers with kernel size three and stride one. The feature aggregator mixes all the representations from the multiple time steps of the feature extractor and computes the contextual representation. The wav2vec model is pre-trained with Libre speech corpus.

Amongst the 1500 channels of the frame-level representations, each channel is weighted by the corresponding scalar weight obtained from the channel attention mechanism. We perform stats pooling that computes mean and standard deviation for each of the three streams to get 3000-dimensional segment level representations. Segment level layers have two fully connected layers, each with 512 units, followed by an output layer with softmax activation. Three individual systems are trained with respect to each feature with dedicated frame level and segment level networks.

The proposed model is trained with three frame-level feature encoders and commonly shared segment-level layers. Proposed feature fusion architecture captures speaker information using two hand-crafted features, MFCC and FDLP coefficients, and speech representations from raw speech samples with the help of a shared segment level layer. B1, B2, and B3 denote MFCC, FDLP, and wav2vec feature encoders. The combined loss is computed by summing the three individual cross-entropy losses with respect to each of the feature representations. The network is trained for three epochs with VoxCeleb1 training data. Stochastic gradient descent with an adam optimizer is used to update the network's parameters. An exponential decay learning rate scheduler is used for optimal training performance. The output activations of the first segment level layer give the 512-dimensional speaker embedding. A dimensionality reduction is performed on these embeddings to 200 dimensions by Linear Discriminant analysis. Later these embeddings are scored with PLDA backend scoring technique [18].

In order to study the noise robustness of each feature and propose a multi-feature integration approach, we evaluate the performance of all the ASV systems on noisy speech data obtained by adding noise to the VoxCeleb1 test data at the various signal to noise ratios (SNR). All the performance results are quoted in the percent Equal Error Rate (EER) and detection cost function (minDCF).

4. RESULTS AND ANALYSIS

Features	EER (%)	minDCF (0.01)	minDCF (0.001)
MFCC (m)	4.942	0.433	0.583
FDLP (f)	5.228	0.433	0.583
wav2vec (w)	3.685	0.397	0.567
Fusion			
Score Fusion (m+f)	4.385	0.412	0.516
Score Fusion (f+w)	3.372	0.377	0.493
Score Fusion (m+w)	3.245	0.340	0.479
Score Fusion (all)	3.234	0.324	0.487

Table 1: Individual ASV System Performance on stand alone features and the corresponding score fusion results.

As shown in the Table.1 among three baseline systems trained with stand-alone features, wav2vec representations are found to be carrying more speaker-specific information when compared with MFCCs and FDLP features. This reflects the potential of deep learning based speech representations than hand-crafted features in speaker discrimination. Also, MFCC features are comparatively performing better than FDLP features in speaker discrimination. The score fusion results as depicted in Table. 1 convey that all these features carry complementary information with respect to each other. It is obvious to expect that the score fusion of hand-crafted MFCC and FDLP features with wav2vec improves the performance of the ASV system.

The experimental results using the proposed multi-feature integration framework are presented in Table. 2. The re-

Feature	EER (%)	minDCF (0.01)	minDCF (0.001)
MFCC (B1)	4.565	0.442	0.551
FDLP (B2)	4.639	0.448	0.540
wav2vec (B3)	3.181	0.366	0.587
Fusion			
Score Fusion (B1+B2)	3.902	0.402	0.514
Score Fusion (B2+B3)	2.953	0.337	0.492
Score Fusion (B1+B3)	3.059	0.322	0.487
Score Fusion (all)	**2.869**	**0.319**	**0.447**

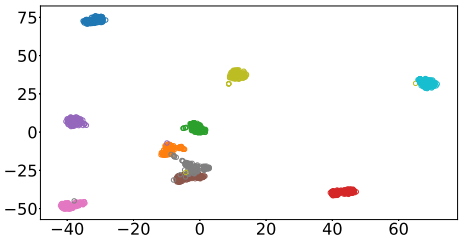
Table 2: System Performance with multi-feature integration and the corresponding score fusion results.

sults clearly showed an improvement over stand-alone systems. For example, the multi-feature integration system with wav2vec features as feature encoder registered a relative improvement of 13% over the stand-alone wav2vec system. Similarly, the proposed system with MFCC and FDLP features encoders attained relative improvements of 7% and 11% over stand-alone MFCC and FDLP systems, respectively. The Same trend is observed in score fusion results as well. When compared with score fusion of three features with stand-alone systems, the score fusion of the proposed system with three feature encoders gains a relative improvement of 11% in performance. And this is achieved with lesser training cost as we need not train exclusively three separate systems. Also, the tSNE plots of speaker embeddings for individual wav2vec and multi-feature integrated wav2vec models are plotted in Fig. 2. This plot shows the low variance and more sparse distribution of speaker clusters with multi-feature integration wav2vec embeddings (Fig. 2(b)) when compared to stand-alone wav2vec based embeddings (Fig. 2(a)).

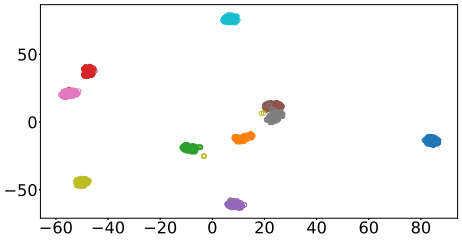
Individual / Feature integrated			
SNR	MFCC	FDLP	wav2vec
clean	4.94 / **4.56**	5.22 / **4.63**	3.68 / **3.18**
30dB	6.31 / **5.75**	6.56 / **6.00**	4.41 / **3.92**
20dB	7.54 / **6.85**	7.68 / **6.92**	5.19 / **4.52**
10dB	10.1 / **9.04**	10.2 / **9.33**	7.02 / **6.06**
0dB	15.9 / **14.6**	16.7 / **15.5**	11.9 / **11.0**

Table 3: Noise robust characteristics of speaker embedding extracted from individual feature and multi-feature integrated speaker models

Extending our experiments to the noise robustness, the results in Table. 3 shows the noise robust characteristics of speaker embedding extracted from the proposed model under various SNR conditions. In all the cases, the speaker embedding extracted from multi-feature integration outperforms the individual feature-based models by significant margins.



(a) individual model



(b) proposed model

Fig. 2: tSNE plots for speaker embeddings extracted from wav2vec representations

5. CONCLUSIONS

These results convey the capability of the proposed multi-feature integration system in harnessing the speaker information from multiple features at the time of training and augmenting the learned information to the given type of features. Having this flexibility of choosing a feature encoder is advantageous at the time of inference. The training complexity is also much reduced due to the shared segment layer while achieving best possible performance exploiting the speaker information present in multiple features. The results also display the noise robust characteristics of speaker embedding extracted by multi-feature integration. Though we have used only three kinds of speech feature representations, the proposed framework is versatile to more number of feature types. As a future scope, the speaker evidence from the phase spectrum along with the magnitude spectrum of the speech signal can be exploited. Also this framework can be extended to other applications of speech such as language and emotion identification to integrate information present in multiple features.

5.1. ACKNOWLEDGMENT

This work was supported by DST National Mission Interdisciplinary Cyber-Physical Systems (NM-ICPS), Technology Innovation Hub on Autonomous Navigation and Data Acquisition Systems: TiHAN Foundations at Indian Institute of Technology (IIT) Hyderabad

6. REFERENCES

[1] Thomas F Quatieri, *Discrete-time speech signal processing: principles and practice*, Pearson Education India, 2006.

[2] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[3] 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, 2010.

[4] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Interspeech*, 2017, pp. 999–1003.

[5] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[6] Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-Jin Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[7] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[8] Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[9] Karthika Vijayan, Pappagari Raghavendra Reddy, and K. Sri Rama Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.

[10] R. Padmanabhan and Hema A. Murthy, "Acoustic feature diversity and speaker verification," in *INTERSPEECH*, 2010.

[11] Nengheng Zheng, Tan Lee, and P. C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Processing Letters*, vol. 14, no. 3, pp. 181–184, 2007.

[12] A. Lawson, P. Vabishchevich, M. Huggins, P. Ardis, B. Battles, and A. Stauffer, "Survey and evaluation of acoustic features for speaker recognition," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5444–5447.

[13] Sriram Ganapathy, Sri Harish Mallidi, and Hynek Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1285–1295, 2014.

[14] Sarthak Yadav and Atul Rai, "Frequency and temporal convolutional attention for text-independent speaker recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6794–6798.

[15] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[16] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[17] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke, "A scalable noisy speech dataset and online subjective test framework," *Proc. Interspeech 2019*, pp. 1816–1820, 2019.

[18] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey*, 2010, vol. 14.