TRANSFER LEARNING USING RAW WAVEFORM SINCNET FOR ROBUST SPEAKER DIARIZATION

Harishchandra Dubey, Abhijeet Sangwan, John H. L. Hansen[+]

Robust Speech Technologies Lab, Center for Robust Speech Systems
The University of Texas at Dallas, Richardson, TX 75080, USA
{Harishchandra.Dubey, Abhijeet.Sangwan, John.Hansen}@utdallas.edu

ABSTRACT

Speaker diarization tells *who spoke and when?* in an audio stream. SincNet is a recently developed novel convolutional neural network (CNN) architecture where the first layer consists of parameterized sinc filters. Unlike conventional CNNs, SincNet take raw speech waveform as input. This paper leverages SincNet in vanilla transfer learning (VTL) setup. Out-domain data is used for training SincNet-VTL to perform frame-level speaker classification. Trained SincNet-VTL is later utilized as feature extractor for in-domain data. We investigated pooling (max, avg) strategies for deriving utterance-level embedding using frame-level features extracted from trained network. These utterance/segment level embedding are adopted as speaker models during clustering stage in diarization pipeline. We compared the proposed SincNet-VTL embedding with baseline i-vector features. We evaluated our approaches on two corpora, CRSS-PLTL and AMI. Results show the efficacy of trained SincNet-VTL for speaker-discriminative embedding even when trained on small amount of data. Proposed features achieved relative DER improvements of 19.12% and 52.07% for CRSS-PLTL and AMI data, respectively over baseline i-vectors.

Index Terms: Speaker Clustering, SincNet, Audio Diarization, Peer-led team learning, Transfer Learning.

1. INTRODUCTION

Speaker Diarization is front-end for multi-subject speech technologies [1]. It provides solution for *who spoke and when?* [2]. In general, it is an unsupervised/semi-supervised system. It consists of sub-systems: (i) speech activity detection (SAD) [3, 4]; (ii) speaker change detection; (iii) clustering; and (iv) re-segmentation where step (iv) is optional. Some approaches combined step (ii) and (iii) into joint segmentation and clustering [5]. Practical applications of speaker diarization [6] include broadcast new analysis, low-latency speaker spotting [7] and behavioral study [8, 2] etc.

State-of-the-art diarization systems use i-vectors in speaker clustering [4]. Recently, neural network embedding (d-vectors) were benchmarked for diarization task. However, most deep neural network based speaker embedding extractor are trained on significantly large amount of data which is not always available [10]. Recently, CNNs were explored for deriving speech representations for a variety of tasks. Such approaches use magnitude spectrum for speech feature learning. The idea of exploring a first layer with parameterized Gaussian filters in a deep neural network was explored for speech recognition [11]. It was trained at frame-level using spectrogram features [11]. Some studies evaluated custom layer consisting of Gabor filters using power-normalized spectrum as input for speech recognition [12]. More recently, using raw waveform for training neural network is an emerging trend. This approach is advantageous as it eliminates the feature extraction pipeline. Learning from time-domain signal showed good results for tasks such as speech recognition [13], emotion identification [14], speaker verification [9] etc.

In this paper, we investigate SincNet for speaker diarization where the first layer consists of sinc filters. Sinc-Layer learns compact band-pass filters suitable for speaker modeling. It is parameterized by cut-off frequencies of these band-pass filters. The gain of sinc filters is learned by later (convolutional and fully connected) layers in SincNet architecture (see Fig. 2). SincNet was developed for speaker recognition in practical scenario where small training data (few seconds/speaker) was available while the test utterances were very short [9]. We leverage efficient SincNet in a vanilla transfer learning (VTL) setup where the SincNet was trained for frame-level speaker recognition on out-domain data and later trained SincNet-VTL was used for extracting speaker embedding from in-domain data (see Fig. 1). We investigated several possibilities for extracting features, namely F1, F2 and F3 that were later pooled to fetch segment-level speaker models. We employed length-normalized SincNet-VTL embedding in a diarization pipeline that uses ground-truth speaker

[+]This project was funded in part by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

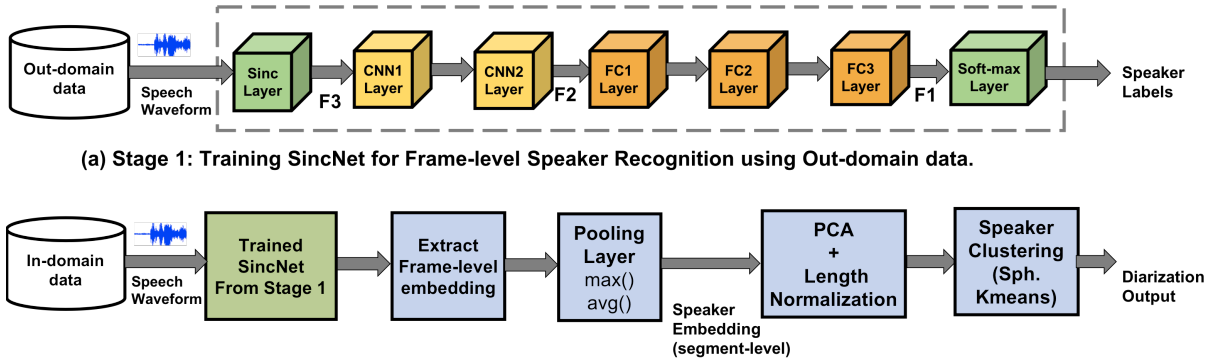

Fig. 1. Proposed SincNet-VTL approach for extracting speaker embedding from time-domain speech. (a) In Stage 1, SincNet is trained for frame-level speaker identification using out-domain data. (b) In Stage 2, we adopt the trained SincNet as feature extractor for in-domain data. We max() or avg() pooled frame-level features to get utterance-level embedding.

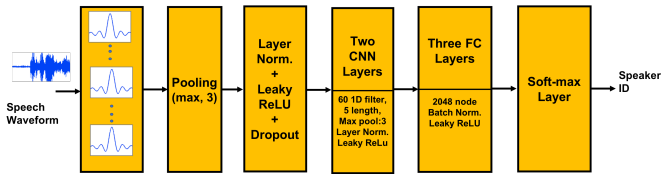

Fig. 2. The architecture of waveform SincNet [9]. Sinc-Layer performs time-domain convolutions on raw speech. Next, two 1D convolutional layers and three fully connected layers filter the input. Final soft-max layer perform speaker classification.

segmentation and spherical K-means clustering.

2. PROPOSED APPROACH

This section explains the proposed approach for SincNet-based vanilla transfer learning (SincNet-VTL) as depicted in Fig. 1. SincNet was trained using out-domain TIMIT data [15]. Trained SincNet was adopted as feature extractor for in-domain data such as CRSS-PLTL and AMI corpora (see Section 3.1).

2.1. SincNet Architecture

Recently, SincNet was developed as an efficient architecture for processing raw speech waveform for speaker recognition [9]. Fig. 2 shows the SincNet architecture that consists of six hidden layers, namely, Sinc-Layer, two 1D convolutional, and three fully connected layers. Sinc-Layer performs sinc-based convolutions on overlapping frames (200ms with 10ms skip rate) of time-domain signal. After the Sinc-Layer, standard CNN pipeline (pooling, batch normalization, ReLU activation, dropout) was employed. As shown in Fig. 1, Sinc-Layer, CNN1 and CNN2 were followed by fully connected layers FC1, FC2 and FC3. Sinc-Layer has 80 sinc filters each with a length of 251 and max pooling over 3. Both CNN1 and CNN2 layers had 60 filters each with length 5 and max pooling over 3. Sinc-Layer, CNN1 and CNN2 employ layer normalization [16] and leaky ReLU activation. Three fully connected layers namely FC1, FC2 and FC3 had same configuration i.e., 2048 nodes, batch normalization [17] and leaky ReLU activation. Final soft-max layer has number of nodes equal to speaker count in the training data. This architecture takes raw speech from 200ms time-windows (frames) with 10ms skip rate and trained for speaker recognition at frame-level.

Sinc-Layer learn the formats and pitch trajectory that facilitate efficient speaker modeling [9] and results in compact representation. Unlike fully connected layers, convolutional ones focus on local regions of the input and extract shift-invariant features that enhances overall recognition performance. Sinc-Layer consists of parameterized sinc functions that act as band-pass filters in spectral domain. Discrete-time sinc filter is given as:

$$h[m, f_1, f_2] = 2f_2 \cdot sinc(2\pi f_2 m) - 2f_1 \cdot sinc(2\pi f_1 m) \quad (1)$$

The $sinc(\cdot)$ functions in above equation is defined as

$$sinc(x) = sin(x)/x. \quad (2)$$

Thus, the Sinc-Layer tries to learn lower and upper cut-off frequencies for filters parameterized by its nodes. For results discussed in this paper, we initialized these with the cutoff frequencies of the Mel filter-bank. Such initialization is preferred as it has more filters in lower frequency spectrum that quantifies speaker characteristics. There are two constraints in Eqn. 1 that need to be satisfied: $f_1 \geq 0$ and $f_2 \geq f_1$. In fact, Eqn. 1 is employed with the following cut-off frequen-

cies:

$$f_1' = |f_1| \\ f_2' = |f_1| + |f_2 - f_1| \quad (3)$$

From above equations, we see that Sinc-Layer tried to learn only the cut-off frequencies. Next, convolutional and fully connected layers learn the gains for each sinc filter by assigning appropriate weights. Passband ripples in sinc filter are mitigated by Hamming windowing that smoothed the abrupt discontinuities. Thus, we have:

$$h_w[m, f_1', f_2'] = h[m, f_1', f_2'] \cdot w_{hamming}[m], \quad (4)$$

where the Hamming window (of length L) is defined as

$$w_{hamming}[m] = 0.54 - 0.46 \cdot cos\left(\frac{2\pi m}{L}\right). \quad (5)$$

The cutoff frequencies of Sinc-Layer are learned jointly with other parameters of SincNet architecture using stochastic gradient descent. SincNet is attractive for speaker modeling due to properties such as fast convergence, compact architecture (few parameters), and computational efficiency (symmetric sinc functions).

2.2. SincNet-VTL for Speaker Modeling

Speaker embedding extracted from trained neural networks are emerging alternatives to i-vectors for speaker modeling. SincNet is a recently developed novel architecture designed for efficient processing of raw waveform [9]. Researchers found SincNet superior to CNN for speaker recognition and verification tasks [9]. We used SincNet trained on out-domain data for vanilla transfer learning (VTL). We propose to leverage out-domain data in speaker diarization through SincNet-VTL approach (see Fig. 1).

We used TIMIT corpus [15] as out-domain data for training SincNet. We ensured text-independent speaker modeling by not including utterances with same text for all speakers, in the training data. Non-speech at the start and end of each utterance was discarded for SincNet training. Time-domain speech signal was divided into 200ms frames with 10ms skip-rate. SincNet was trained using raw speech waveform for frame-level speaker recognition. Sinc-layer parameters were initialized with Mel-scale cutoff frequencies while rest of the network was initialized with Glorot scheme [18]. Final soft-max layer implements frame-level speaker classification. The complete network was trained jointly using RMSprop optimizer with learning rate 0.001. We trained it for 360 epochs with batch size of 64. Trained SincNet-VTL has 462 nodes in output layer corresponding to speakers in training data. We tuned network hyper-parameters on TIMIT corpus. During embedding extraction on CRSS-PLTL corpus there were some segments that lasts for less than 200ms. We repeated those segments until it becomes a segment of 1s for getting speaker embedding. Since SincNet-VTL was trained on 200ms windows with 10ms skip-rate, we needed at-least 200ms for doing a forward pass on trained SincNet-VTL. We propose pooling frame-level features extracted using trained SincNet-VTL for getting segment-level embedding (see Fig. 1b).

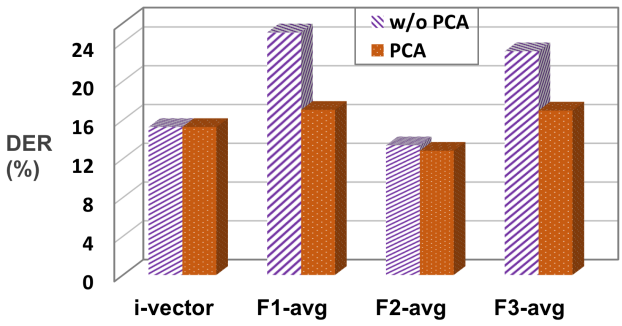

Fig. 3. PLTL data: Comparing i-vector with average pooled F1, F2 and F3 embeddings. w/o PCA means without PCA based dimension reduction.

3. EXPERIMENTS & RESULTS

3.1. CRSS-PLTL Corpus

In collaboration with Student Success Center at UT Dallas, we collected the CRSS-PLTL corpus [2, 19]. It contains multi-stream audio recordings from five PLTL teams over 11 week each, thus leading to 55 sessions. These five teams were chosen from an undergraduate chemistry course. Each PLTL session lasted for approximately 80 minutes and constitute discussions between 6-8 students plus a peer-leader. Peer leader guides the group to arrive at correct solutions without explicitly telling the solution.

During PLTL sessions, each participant wore a LENA device (with not-so-close microphone) for collecting naturalistic audio [20, 21, 22]. In this manner, we collected multi-stream audio for each session (number of streams was same as total participants). The salient features of this data are: (i) many segments with overlapped-speech; (ii) short conversational-turns; (iii) multiple noise-sources; and (iv) reverberation. These factors made PLTL speaker diarization challenging. In this paper, we choose the channel corresponding to PLTL leader for single-channel diarization evaluation. This evaluation set has 8 speakers and lasted for about 80 minutes. It is important to note that many speaker turns lasted for less than or equal to 1 second.

3.2. AMI Corpus

Augmented Multi-party Interaction (AMI) corpus provides speaker annotated multi-modal data from meeting scenarios. In this paper, we choose 6 meetings from AMI corpus as evaluation set. These six meetings have four speakers each. We

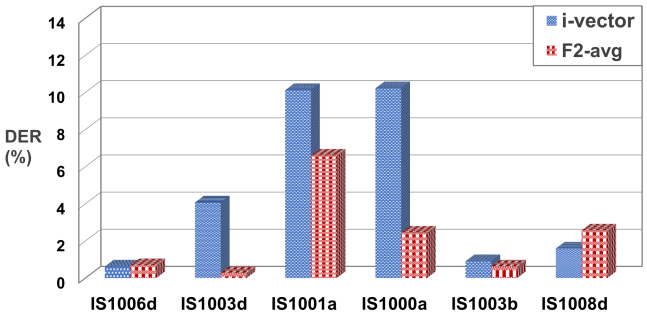

Fig. 4. AMI: 6 meetings. F2-avg with PCA (51 dim.) shows significant improvements over i-vector with PCA (51 dim.).

used *mixed headset audio* for experiments reported in this paper. Our AMI evaluation set consists of sessions: IS1006d (31 min.), IS1003d (36 min.), IS1001a (16 min.), IS1000a (27 min.), IS1003b (27 min.) and IS1008d (25 min.).

3.3. Evaluations

Diarization error rate (DER) was used for scoring the systems with respect to ground-truth annotations. It was introduced in the NIST Rich Transcription Spring 2003 evaluation (RT-03S). It is defined as the total percentage of reference time that is not correctly attributed to a speaker. Mathematically, DER is given as:

$$\text{DER} = \frac{\Phi_{fa} + \Phi_{miss} + \Phi_{spk}}{\Phi_{total}}, \tag{6}$$

where Φ_{total} is the total time of all reference segments, Φ_{fa} is the system speaker-time not attributed to the reference speaker, Φ_{miss} is the total reference speaker-time not attributed to a system speaker, and Φ_{spk} is the total reference speaker-time attributed to a wrong speaker. Unlike NIST RT evaluations [23], no forgiveness collar was allowed during scoring for results presented in this paper. We adopted the NIST md-eval scoring script (version-22) for DER computations.

We kept all audio data at 16 kHz for experiments reported in this paper. We trained an i-vector extractor on TIMIT using ground-truth SAD. Since main focus of this paper is to develop a speaker model for diarization, we used ground-truth speaker segmentation information. In this paper, we adopted 75-dimensional (dim.) i-vector as many segments in PLTL were approximately 1s duration (or shorter). SincNet speaker embeddings has dimensions: F1 (462), F2 (2048), F3 (6420). For all our experiments reported here, we perform length-normalization of i-vectors/embeddings followed by spherical K-means clustering with cosine similarity. For some experiments, we employed principal component analysis (PCA) for dimension reduction to 51. Table 1 shows the effect of PCA (51 dimension) on DER (%) for three features:

Table 1. CRSS-PLTL data: Effect of PCA (51 components) on DER (%) for i-vector, F2-avg and F2-max features.

	i-vector	**F2-avg**	**F2-max**
w/o PCA	15.26	13.37	43.55
PCA	15.26	**12.81**	14.36

i-vectors, F2-avg and F2-max where the latter two are average and max pooled version of frame-level F2 embeddings from trained SincNet-VTL network (see Fig. 1). Fig. 3 shows the comparison of i-vector with average pooled F1, F2 and F3 embeddings with PCA on PLTL data. Fig. 4 shows DER for 6 meetings of AMI corpus (see Section 3.2) using i-vector baseline and best proposed feature, i.e., F2-avg.

4. DISCUSSIONS & CONCLUSIONS

We employed principal component analysis (PCA) for dimension reduction of speaker embedding and i-vectors. We choose PCA with 51 components for both CRSS-PLTL and AMI evaluation sets. Since comparative studies in this paper were focused on speaker modeling, our diarization pipeline consists of ground-truth speaker segmentation and uses spherical K-means clustering with cosine similarity. SincNet-VTL embedding (F1/F2/F3) or i-vectors were extracted from all segments for speaker modeling. We always perform length normalization of speaker features just before clustering. Some experiments had PCA-based dimension reduction prior to length-normalization.

Looking at Table 1, we see that i-vector did not get DER improvements from PCA as i-vectors were already lower dimensional. We see F2-max embedding have benefited the most with PCA. Even if F2-avg has got relatively small reduction in DER with PCA as compared to F2-max, we get the best DER using PCA on F2-avg embedding. After this point, we stick to average pooling as it was better than max pooling for all the three embedding. Fig. 3 shows that F2-avg is best feature for speaker diarization leading to absolute and relative DER improvements of 2.45% and 19.12%, respectively with respect to i-vector baseline. These comparisons were done on PLTL evaluation set as it is our target domain. Fig. 4 shows comparison of F2-avg with i-vector features for AMI data. Proposed F2-avg embedding gave significant DER (%) improvement as compared to i-vector baseline on AMI data. On average, F2-avg leads to absolute and relative DER improvements of 2.39 % and 52.06%, respectively over i-vectors. In this paper, two type of pooling operations were performed on frame-level features to get segment-level embedding : max() and avg(). While max() pooling pick maximum value along each feature dimensions, avg() pooling averages along each dimension. As the results showed in last section, avg() pooling performs better than max() for all three types of proposed speaker embedding.

5. REFERENCES

[1] H. Dubey, A. Sangwan, and J. H. L. Hansen, "Robust speaker clustering using mixtures of von mises-fisher distributions for naturalistic audio streams," in *ISCA INTERSPEECH*, 2018, pp. 3603–3607.

[2] ——, "Using speech technology for quantifying behavioral characteristics in peer-led team learning sessions," *Computer Speech & Language*, vol. 46, pp. 343–366, 2017.

[3] ——, "Leveraging Frequency-Dependent Kernel and DIP-Based Clustering for Robust Speech Activity Detection in Naturalistic Audio Streams," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2056–2071, 2018.

[4] ——, "Robust feature clustering for unsupervised speech activity detection," in *IEEE ICASSP*, 2018, pp. 2726–2730.

[5] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Trans. on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1086–1099, 2018.

[6] H. Dubey, A. Sangwan, and J. H. L. Hansen, "A robust diarization system for measuring dominance in peer-led team learning groups," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 319–323.

[7] J. Patino, R. Yin, H. Delgado, H. Bredin, A. Komaty, G. Wisniewski, C. Barras, N. Evans, and S. Marcel, "Low-latency speaker spotting with online diarization and detection," in *ISCA Odyssey*, 2018, pp. 140–146.

[8] J. H. L. Hansen, A. Sangwan, A. Ziaei, H. Dubey, L. Kaushik, and C. Yu, "Prof-Life-Log: Monitoring and assessment of human speech and acoustics using daily naturalistic audio streams," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3010–3010, 2016.

[9] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *IEEE SLT*, 2018.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," *IEEE ICASSP*, 2018.

[11] H. Seki, K. Yamamoto, and S. Nakagawa, "A deep neural network integrated with filterbank learning for speech recognition," in *IEEE ICASSP*, 2017, pp. 5480–5484.

[12] S.-Y. Chang and N. Morgan, "Robust CNN-based speech recognition with Gabor filter kernels," in *ISCA INTERSPEECH*, 2014.

[13] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *ISCA INTERSPEECH*, 2015.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE ICASSP*, 2016, pp. 5200–5204.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus," in *CDROM*, 1993.

[16] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[18] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[19] J. H. L. Hansen, J. Alberte, N. Jones, H. Dubey, and A. Sangwan, "Multi-stream audio analysis for knowledge extraction and understanding of small-group interactions in peer-led team learning," *Seventh Annual Conference Peer-Led Team Learning International Society, the University of Texas at Dallas, Richardson, TX, USA*, pp. 1–1, 2018.

[20] H. Dubey, L. Kaushik, A. Sangwan, and J. H. L. Hansen, "A speaker diarization system for studying peer-led team learning groups," in *ISCA INTERSPEECH*, 2016, pp. 2180–2184.

[21] J. H. L. Hansen, H. Dubey, A. Sangwan, L. Kaushik, and V. Kothapally, "UTDallas-PLTL: Advancing multistream speech processing for interaction assessment in peer-led team learning," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1869–1869, 2018.

[22] J. H. L. Hansen, H. Dubey, and A. Sangwan, "CRSS-LDNN: Long-duration naturalistic noise corpus containing multi-layer noise recordings for robust speech processing," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1797–1797, 2018.

[23] "The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan." [Online]. Available: https://web.archive.org/web/20100606092041if_/http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf