H-VECTORS: UTTERANCE-LEVEL SPEAKER EMBEDDING USING A HIERARCHICAL ATTENTION MODEL

Yanpei Shi[*]*, Qiang Huang*[*]*, Thomas Hain*

Speech and Hearing Research Group
Department of Computer Science, University of Sheffield
{YShi30, qiang.huang, t.hain}@sheffield.ac.uk

ABSTRACT

In this paper, a hierarchical attention network is proposed to generate utterance-level embeddings (H-vectors) for speaker identification and verification. Since different parts of an utterance may have different contributions to speaker identities, the use of hierarchical structure aims to learn speaker related information locally and globally. In the proposed approach, frame-level encoder and attention are applied on segments of an input utterance and generate individual segment vectors. Then, segment level attention is applied on the segment vectors to construct an utterance representation. To evaluate the effectiveness of the proposed approach, the data of the NIST SRE2008 Part1 is used for training, and two datasets, the Switchboard Cellular (Part1) and the CallHome American English Speech, are used to evaluate the quality of extracted utterance embeddings on speaker identification and verification tasks. In comparison with two baselines, X-vectors and X-vectors+Attention, the obtained results show that the use of H-vectors can achieve a significantly better performance. Furthermore, the learned utterance-level embeddings are more discriminative than the two baselines when mapped into a 2D space using t-SNE.

Index Terms— Speaker Embeddings, Speaker Identification, Hierarchical Attention, X-vectors, Attention Mechanism

1. INTRODUCTION

The generation of compact representation used to distinguish speakers has been an attractive topic and widely used in some studies, such as speaker identification [17], verification [19, 14, 11], detection [13], segmentation [7, 23], and speaker dependent speech enhancement [2, 6].

To extract a general speaker representation, Najim et al. [5] defined a "total variability space" containing the speaker and channel variabilities simultaneously, and then extracted the speaker factors by decomposing feature space into subspace corresponding to sound factors including speaker and channel effects. With the rapid development of deep learning technologies, some architectures using deep neural networks (DNN) have been developed for general speaker representation [22, 20]. In [22], Variani et al. introduced the d-vector approach using the LSTM and averaging over the activations of the last hidden layer for all frame-level features. David et al. [20] used a five-layer DNN with taking into account a small temporal context and statistics pooling. To further improve the embedding quality, attention mechanisms have been used in some recent studies [24, 26]. Wang, et al. [24] designed an attentive X-vector where a self-attention layer was added before a statistic pooling layer to weight each frame vector.

However, how to highlight the importance of different parts of an input utterance is underdeveloped . For this issue, a hierarchical attention mechanism is employed in this paper. This is inspired by Yang's work [25] in document classification, where it claimed that not all parts of a document are equally relevant for answering a query and attention models were thus applied to both word and sentence level feature vectors via a hierarchical network. In the proposed approach, an utterance can be viewed as a document, and its divided segments and acoustic frames are treated as sentences and words, respectively. An attention mechanism is then used at both frame level and segment level. The utterance embedding can be constructed by first building representations of segments from frames and then aggregating those into an utterance representation. The use of this hierarchical attention network (HAN) can offer a way to obtain a discriminative utterance-level embedding by explicitly weight target relevant features.

The rest of the paper is organized as follow: Section 2 presents the architecture of our approach. Section 3 depicts the used data, experimental setup, and the baselines to be compared. The obtained results are shown in Section 4, and a conclusion is finally drawn in Section 5.

2. MODEL ARCHITECTURE

Figure 1 shows the architecture of a hierarchical attention network. The network consists of several parts: a frame-level encoder and attention layer, a segment-level encoder and attention layer, and two fully connect layers. Given input acoustic frame vectors, the proposed model generates an utterance-

*The first and second author contribute equally to this paper

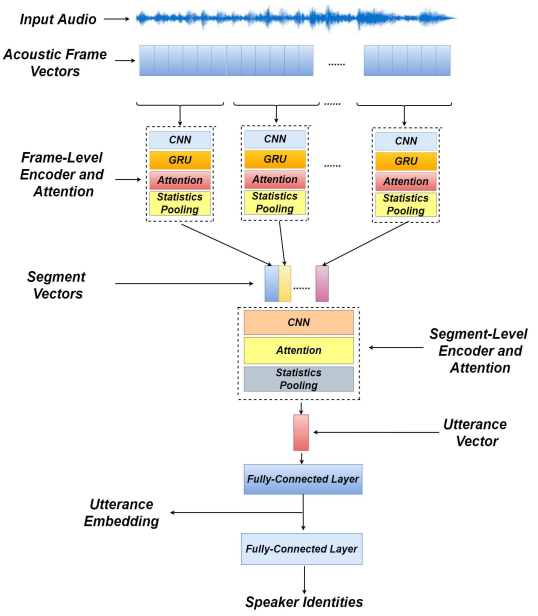


Fig. 1. Architecture of Hierarchical Attention Network.

level embedding, by which a classifier is trained to perform speaker identification or verification. The details of each part will be introduced in the following subsections.

2.1. Frame-Level Encoder and Attention

Assume that an utterance is divided into N segments: $\mathbf{S} \in \mathcal{R}^{MN\times L} = \{\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_N\}$ with a fixed-length window. Each segment $\mathbf{S}_i \in \mathcal{R}^{M\times L} = \{\mathbf{x}_{i1}, \mathbf{x}_{i,2}, \cdots, \mathbf{x}_{i,M}\}$ contains M L-dimensional acoustic frame vectors $\boldsymbol{x}_{i,t} \in \mathcal{R}^{1\times L}$, where i denotes the ith segment, t denotes the tth frame, $i \in \{1, \cdots N\}, t \in \{1, \ldots, M\}$.

In the frame-level encoder, a one-dimensional CNN is used on each segment, and followed by a bidirectional GRU [3] in order to get information from both directions of acoustic frames and contextual information.

$$\begin{aligned}\boldsymbol{S}_i^{'} &= \mathrm{CNN}(\boldsymbol{S}_i)\\ \overrightarrow{\boldsymbol{h}}_i &= \overrightarrow{\mathrm{GRU}}(\boldsymbol{S}_i^{'})\\ \overleftarrow{\boldsymbol{h}}_i &= \overleftarrow{\mathrm{GRU}}(\boldsymbol{S}_i^{'})\end{aligned}$$

The output of a frame-level encoder $\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}}_i, \overleftarrow{\boldsymbol{h}}_i] \in \mathcal{R}^{M\times E} = \{\boldsymbol{h}_{i,1}, \boldsymbol{h}_{i,2}, \cdots, \boldsymbol{h}_{i,M}\}$ contains the information of the segment $\boldsymbol{S}_i$.

In the frame-level attention layer, a two-layer MLP is first used to convert $\boldsymbol{h}_i$ into score vector $\boldsymbol{z}_i$, by which a normalised importance weight vector $\boldsymbol{\alpha}_i$ can be computed via a softmax function [25].

$$\alpha_{i,t} = \frac{\exp(z_{i,t})}{\sum_{t=0}^{M}\exp(z_{i,t})} \tag{1}$$

$$z_{i,t} = \mathrm{Relu}(\boldsymbol{h}_{i,t}\boldsymbol{W}_{i,0} + \boldsymbol{b}_{i,0})\boldsymbol{W}_{i,1} \quad , \tag{2}$$

where $z_{i,t}$ and $\alpha_{i,t}$ are a scalar score and normalized score for each time step t respectively. $\boldsymbol{W}_{i,0} \in \mathcal{R}^{E\times E}$, $\boldsymbol{b}_{i,0} \in \mathcal{R}^{1\times E}$ and $\boldsymbol{W}_{i,1} \in \mathcal{R}^{E\times 1}$ are the parameters of a two-layer MLP. These parameters are shared when processing N segments. A weighted output of frame-level encoder is computed by

$$\boldsymbol{A}_{i,t} = \alpha_{i,t}\boldsymbol{h}_{i,t} \tag{3}$$

Following [20], a statistics pooling is applied on $\boldsymbol{A}_i$ to compute its mean vector ($\boldsymbol{\mu}_i$) and std ($\boldsymbol{\sigma}_i$) vector over t. A segment vector $\boldsymbol{V}_{S_i}$ is then obtained by concatenating the two vectors:

$$\boldsymbol{V}_{S_i} = \mathrm{concatenate}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \tag{4}$$

2.2. Segment Level Encoder and Attention

For the segment-level encoder and attention, the same steps used in frame-level encoder and attention are followed except for a bi-directional GRU layer, as the omission of the GRU layer can well accelerate training when processing a large number of samples.

The output of the frame level encoder and attention is $\boldsymbol{V}_S \in \mathcal{R}^{N\times E} = \{\boldsymbol{V}_{S_1}, \boldsymbol{V}_{S_2}, \cdots, \boldsymbol{V}_{S_N}\}$. The weight vector $\boldsymbol{\alpha}^s \in \mathcal{R}^{N\times 1} = \{\alpha_1^s, \alpha_2^s, \cdots, \alpha_N^s\}$ of segment level attention can be computed as follows [15]:

$$\begin{aligned}\alpha_i^s &= \frac{\exp(z_i^s)}{\sum_{i=0}^{N}\exp(z_i^s)}\\ z_i^s &= \mathrm{Relu}(\boldsymbol{V}_{S_i}\boldsymbol{W}_{n,0} + \boldsymbol{b}_{n,0})\boldsymbol{W}_{n,1} \quad ,\end{aligned} \tag{5}$$

where z_i^s and α_i^s are a scalar score and normalized score for each segment vector $\boldsymbol{V}_{S_i}$ respectively. $\boldsymbol{W}_{n,0} \in \mathcal{R}^{E\times E}$, $\boldsymbol{b}_{n,0} \in \mathcal{R}^{1\times E}$ and $\boldsymbol{W}_{n,1} \in \mathcal{R}^{E\times 1}$ are the parameters of a two-layer MLP. A vector is generated using a statistics pooling over all weighted segments:

$$\begin{aligned}\boldsymbol{\mu}_U &= mean(\sum_i \alpha_i^s \boldsymbol{S}_i)\\ \boldsymbol{\sigma}_U &= std(\sum_i \alpha_i^s \boldsymbol{S}_i)\\ \boldsymbol{V}_U &= \mathrm{concatenate}(\boldsymbol{\mu}_U, \boldsymbol{\sigma}_U)\end{aligned} \tag{6}$$

The final speaker identity classifier is constructed using a two-layer MLP with $\boldsymbol{V}_U$ being its input. As shown in figure 1, the output of the first fully connected layer can be used as the final utterance embedding, represented by $\boldsymbol{E}mb_U$.

3. EXPERIMENT

3.1. Data

Three datasets, NIST SRE 2008 part1 (SRE08), CallHome American English Speech (CHE), and Switchboard Cellular Part 1 (SWBC), are used in this paper to train the proposed model and evaluate utterance embedding performance. SRE08 indicates the 2008 NIST speaker recognition evaluation test set [8], which contains multilingual telephone speech and English interview speech. In this work, Part1

Dataset	Type	#Speaker	Size (hour)	#Utterance (1s)	#Utterance (3s)
SRE08	Telephone+Interview	1336	640	3,528,326	1,176,453
CHE	Telephone	120	60	252,224	84,460
SWBC	Telephone	254	130	1,008,901	336,417

Table 1. Details of three telephone speech datasets: Part1 of Sre2008 (SRE08), CallHome(CHE), and Switchboard(SWBC).

of SRE2008, containing about 640-hour speech and 1336 distinct speakers, is selected in our experiments.

SWBC [4] contains 130 hours telephone speech, totally 254 speakers (129 male and 125 female) under various environment conditions (indoors, outdoors and moving vehicles). The stereo speech signals are split into two monos, and both of them are used in experiments. CHE [1] contains 120 telephone conversations speech between native English speakers (totally 120 speakers). Among all of the calls, 90 of them are placed to various locations outside North America. In this dataset, speech from the left channel is used, as the labels of speakers in the right channels is unavailable. In our experiments, SRE08 is used to train the proposed model, by which Utterance-level embeddings can be then generated using CHE and SWBC.

3.2. Experiment Setup

In this work, after removing unvoiced signals using energy based VAD [16], fixed length sliding windows (one second or thre seconds) with half-size shift is employed to divide speech streams into short segments. Each segment is viewed as an utterance independently. The total number of utterances of the three datasets are listed in Table 1. Each utterance is then split into 10 equal-length fragments without overlap. Each fragment is further segmented into frames using a 25ms sliding window with a 10ms shift. All frames are converted into 20-dimensional MFCC feature vectors. Similar to [25], to build a hierarchical structure, each utterance, fragment and frame vector obtained here are viewed as a document, sentence and word, respectively.

To evaluate the utterance-level embeddings, speaker identification and verification are conducted using the utterance-level embeddings generated on CHE and SWBC. Instead of directly processing on the embeddings, PLDA back-end [18] is applied on the embeddings to reduce the dimension to 300.

Both the SWBC and the CHE datasets are randomly split into training and test data with 9:1 ratio for speaker identification. For a speaker verification task, in SWBC, there are 50 speakers in the enrolment set and 120 speakers in the evaluation set, with 10 utterances for each speaker. In the CHE, there are 30 speakers in the enrolment set and 60 speakers in the evaluation set. Each speaker has 10 utterances.

In order to compare the proposed approach with other speaker embedding systems, two baselines are built using the methods developed in previous studies. The first baseline ("X-Vectors") is based on a TDNN architecture [20]. It is now widely used for speaker recognition and is effective in speaker embedding extraction. The second baseline ("X-Vectors+Attention") is made by combining a global attention mechanism with X-vectors. [24, 26]. For evaluation, in our speaker identification task, prediction accuracy is reported in this work. In the speaker verification task, the equal error rate (EER) is reported. Moreover, to show the quality of the learned utterance-level embeddings, t-SNE [12] is used to visualize their distributions after being projected in a 2-dimensional space.

Level	Model	Input	Output
Frame-Level	CNN	(30,20,1)	(30,1,512)
	Bi-GRU	(30,512)	(30,1024)
	Attention	(30,1024)	(30,1024)
	Statistics Pooling	(30,1024)	(1,2048)
Segment-Level	CNN	(10,2048,1)	(10,1,1500)
	Attention	(10,1500)	(10,1500)
	Statistics Pooling	(10,1500)	(1,3000)
Utterance-Level	DNN (512)	(1,3000)	(1,512)
	DNN (512)	(1,512)	(1,512)

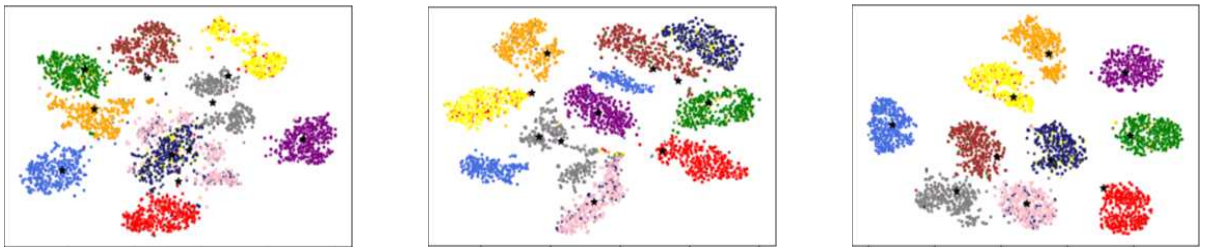
Table 2. Architecture of the proposed approach

Table 2 shows the configuration of the proposed architecture. It also contains batch normalization [9] and dropout [21] layers, where the dropout rate is set to 0.2. Adam optimiser [10] is used for all experiments with $\beta_1 = 0.95$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The initial learning rate is 10^{-4}.

4. RESULTS

Table 3 shows the prediction accuracy on the test data of SRE08 using the proposed approach and two baselines. Two different utterance lengths, 1 second and 3 seconds, are used in the experiments, respectively. The use of the H-vectors shows higher accuracy when using either 1-second or 3-second input length than the two baselines. When the length of input utterances is one second, the accuracy obtained using the H-vectors can reach 94.5%, with 4.4% improvement over X-vectors and 2.4% improvement over X-vectors+Attention, respectively. When the length of input utterances is three seconds, the accuracy obtained using the H-vectors can reach 98.5%, with about 3% improvement over X-vectors and about 2% improvement over X-vectors+Attention. The proposed approach is more robust than the two baselines when processed utterances are short. In addition, the accuracies obtained using 3-second utterances are better than those using 1-second utterances. This probably means a longer utterance may contain more information relevant to a target speaker than short ones.

To evaluate the quality of embeddings extracted using the



(a) X-vector (b) X-vector+Attention (c) H-vector

Fig. 2. Embedding visualization using t-SNE. Each color represents a speaker, and each point indicates an utterance.

Utterance Length	Model	Accuracy %
1 Second	X-vector	90.1
	X-vector+Attention	92.1
	H-vector	94.5
3 Seconds	X-vector	95.2
	X-vector+attention	96.7
	H-vector	98.5

Table 3. Identification accuracy on the test data of SRE08 when the utterance length is 1s or 3s.

Utterance Length	Model	Accuracy %	EER %
1 Second	X-vector	84.8	1.94
	X-vector+Attention	87.5	1.61
	H-vector	89.1	1.44
3 Seconds	X-vector	89.4	1.46
	X-vector+attention	91.0	1.21
	H-vector	92.8	1.08

Table 4. Identification accuracy and Equal Error Rate (EER) on CHE dataset when the utterance length is 1s or 3s.

proposed approach, two additional datasets are employed in our experiments. Table 4 and Table 5 show the identification accuracy and verification EER when using the embeddings extracted on the SWBC and the CHE dataset, respectively. On the two datasets, the H-vectors consistently outperforms the two baselines whether the length of utterances is one second or three seconds.

Since the model is trained on the SRE08 corpus, the identification performances on its test data are clearly better than those on the other two datasets. As the SWBC dataset contains a wide range of environment conditions (indoors, outdoors and moving vehicles), both its identification and verification performances are relatively worse than those obtained on the CHE dataset.

To further test the quality of extracted utterance-level embeddings, t-SNE [12] is used to visualise the distribution of embeddings by projecting these high-dimensional vectors into a 2D space. In the SWBC dataset, 10 speakers are selected and 500 three-second segment are randomly sampled for each speaker. Figure 2 (a), (b), and (c) show the distribution of selected samples of 10 speakers after using X-vectors, X-vectors+Attention, and H-vectors, respectively. Each color represents a single distinct speaker and each point represents an utterance. The black mark represents the center point of

Utterance Length	Model	Accuracy %	EER %
1 Second	X-vector	78.2	2.23
	X-vector+Attention	81.0	2.05
	H-vector	83.7	1.92
3 Seconds	X-vector	81.3	2.01
	X-vector+attention	84.0	1.82
	H-vector	86.2	1.69

Table 5. Identification accuracy and Equal Error Rate (EER) on SWBC dataset when the utterance length is 1s or 3s.

each speaker class. Figure 2(a) shows the distribution of the embeddings obtained by X-vectors. It is clear that, in this figure, some samples from different speakers are not well discriminated as there are overlaps between speaker classes. Due to the use of an attention mechanism in X-vectors+Attention, figure 2(b) shows a better sample distribution than figure 2(a). However, some samples of a speaker labelled by a blue colour are not well clustered. In figure 2(c), the embedding obtained by H-vectors performs a better separation property than the other two baselines.

5. CONCLUSION AND FUTURE WORK

In this paper, a hierarchical attention network was proposed utterance-level embedding extraction. Inspired by the hierarchical structure of a document made by words and sentences, each utterance is viewed as a document, segments and frame vectors are treated as sentences and words, respectively. The use of attention mechanisms at frame and segment levels provides a way to search for the information relevant to target locally and globally, and thus obtained better utterance level embeddings, including better performance on speaker identification and verification tasks using the extracted embeddings. Moreover, the obtained utterance-level embeddings are more discriminative than the use of X-vectors and X-vectors+Attention.

In the future work, different kinds of acoustic features such as filter-bank and Mel-spectrogram will be investigated and tested on some large datasets, such as Voxceleb1 and 2.

Acknowledgement

This work was in part supported by Innovate UK Grant number 104264 MAUDIE and Huawei Innovation Research Program (HIRP) number X/159898-11.

6. REFERENCES

[1] ALEXANDRA CANAVAN, DAVID GRAFF, G. Z. Callhome american english speech. https://catalog.ldc.upenn.edu/LDC97S42, 2001.

[2] CHUANG, F.-K., WANG, S.-S., HUNG, J.-W., TSAO, Y., AND FANG, S.-H. Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement. *Interspeech* (2019), 3173–3177.

[3] CHUNG, J., GULCEHRE, C., CHO, K., AND BENGIO, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS* (2014).

[4] DAVID GRAFF, KEVIN WALKER, D. M. Switchboard cellular part 1 audio. https://catalog.ldc.upenn.edu/LDC2001S13, 2001.

[5] DEHAK, N., KENNY, P. J., DEHAK, R., DUMOUCHEL, P., AND OUELLET, P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* (2010), 788–798.

[6] GAO, T., DU, J., XU, L., LIU, C., DAI, L.-R., AND LEE, C.-H. A unified speaker-dependent speech separation and enhancement system based on deep neural networks. In *ChinaSIP* (2015), IEEE.

[7] GARCIA-ROMERO, D., SNYDER, D., SELL, G., POVEY, D., AND MCCREE, A. Speaker diarization using deep neural network embeddings. In *ICASSP* (2017), IEEE.

[8] GROUP, N. M. I. 2008 nist speaker recognition evaluation training set part 1. https://catalog.ldc.upenn.edu/LDC2011S05, 2011.

[9] IOFFE, S., AND SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML* (2015), pp. 448–456.

[10] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2014).

[11] LE, N., AND ODOBEZ, J.-M. Robust and discriminative speaker embedding via intra-class distance variance regularization. In *Interspeech* (2018), pp. 2257–2261.

[12] MAATEN, L. V. D., AND HINTON, G. Visualizing data using t-sne. *JMLR* (2008), 2579–2605.

[13] MCLAREN, M., CASTÁN, D., NANDWANA, M. K., FERRER, L., AND YILMAZ, E. How to train your speaker embeddings extractor. In *Odyssey* (2018).

[14] NOVOSELOV, S., SHULIPA, A., KREMNEV, I., KOZLOV, A., AND SHCHEMELININ, V. On deep speaker embeddings for text-independent speaker recognition. In *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop* (2018), pp. 378–385.

[15] PAN, Y., MIRHEIDARI, B., REUBER, M., VENNERI, A., BLACKBURN, D., AND CHRISTENSEN, H. Automatic hierarchical attention neural network for detecting ad. In *Interspeech* (2019).

[16] PANG, J. Spectrum energy based voice activity detection. *CCWC* (2017), 1–5.

[17] PARK, H., CHO, S., PARK, K., KIM, N., AND PARK, J. Training utterance-level embedding networks for speaker identification and verification. In *Interspeech* (2018).

[18] SALMUN, I., OPHER, I., AND LAPIDOT, I. On the use of plda i-vector scoring for clustering short segments. In *Odyssey* (2016), pp. 407–414.

[19] SNYDER, D., GARCIA-ROMERO, D., POVEY, D., AND KHUDANPUR, S. Deep neural network embeddings for text-independent speaker verification. In *Interspeech* (2017).

[20] SNYDER, D., GARCIA-ROMERO, D., SELL, G., POVEY, D., AND KHUDANPUR, S. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP* (2018), IEEE, pp. 5329–5333.

[21] SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I., AND SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. *JMLR* (2014), 1929–1958.

[22] VARIANI, E., LEI, X., MCDERMOTT, E., MORENO, I. L., AND GONZALEZ-DOMINGUEZ, J. Deep neural networks for small footprint text-dependent speaker verification. In *ICASSP* (2014), IEEE.

[23] WANG, Q., DOWNEY, C., WAN, L., MANSFIELD, P. A., AND MORENO, I. L. Speaker diarization with lstm. In *ICASSP* (2018), IEEE, pp. 5239–5243.

[24] WANG, Q., OKABE, K., LEE, K. A., YAMAMOTO, H., AND KOSHINAKA, T. Attention mechanism in speaker recognition: What does it learn in deep speaker embedding? In *SLT* (2018), IEEE, pp. 1052–1059.

[25] YANG, Z., YANG, D., DYER, C., HE, X., SMOLA, A., AND HOVY, E. Hierarchical attention networks for document classification. In *NAACL* (2016), pp. 1480–1489.

[26] ZHU, Y., KO, T., SNYDER, D., MAK, B., AND POVEY, D. Self-attentive speaker embeddings for text-independent speaker verification. In *Interspeech* (2018), pp. 3573–3577.