

DOUBLE MULTI-HEAD ATTENTION FOR SPEAKER VERIFICATION

Miquel India, Pooyan Safari, Javier Hernando

TALP Research Center, Department of Signal Theory and Communications,
Universitat Politècnica de Catalunya, Barcelona, Spain
{miquel.angel.india, javier.hernando}@upc.edu, pooyan.safari@tsc.upc.edu

ABSTRACT

Most state-of-the-art Deep Learning systems for speaker verification are based on speaker embedding extractors. These architectures are commonly composed of a feature extractor front-end together with a pooling layer to encode variable-length utterances into fixed-length speaker vectors. In this paper we present Double Multi-Head Attention pooling, which extends our previous approach based on Self Multi-Head Attention. An additional self attention layer is added to the pooling layer that summarizes the context vectors produced by Multi-Head Attention into a unique speaker representation. This method enhances the pooling mechanism by giving weights to the information captured for each head and it results in creating more discriminative speaker embeddings. We have evaluated our approach with the VoxCeleb2 dataset. Our results show 6.09% and 5.23% relative improvement in terms of EER compared to Self Attention pooling and Self Multi-Head Attention, respectively. According to the obtained results, Double Multi-Head Attention has shown to be an excellent approach to efficiently select the most relevant features captured by the CNN-based front-ends from the speech signal.

Index Terms— self multi-head attention, double attention, speaker recognition, speaker verification

1 Introduction

Speaker verification aims to determine whether a pair of audios corresponds to the same speaker. Given speech signals, speaker verification systems are able to extract speaker identity patterns from the characteristics of the voice. These patterns can be both statistically modelled or encoded into discriminative speaker representations. Over the last few years, researchers have put huge effort on encoding these speaker characteristics into more discriminative speaker vectors. Current state-of-the-art speaker verification systems are based on Deep Learning (DL) approaches. These architectures are commonly trained as speaker classifiers in order to be used as speaker embedding extractors. Speaker embeddings are fixed-length vectors extracted from some of the last layers of these Deep Neural Networks (DNNs) [1]. The most known representation is the x-vector [2], which has become state-of-the-art for speaker recognition and has also been used for other tasks such as language and emotion recognition [3, 4].

Most of the recent network architectures used for speaker embedding extraction are composed of a front-end feature extractor, a pooling layer, and a set of Fully Connected (FC) layers. Lately, there have been several architectures proposed to encode audio utterances into speaker embeddings for different choices of network

inputs, such as [5, 6, 7, 8, 9]. Using Mel-Frequency Cepstral Coefficient (MFCC) features, Time Delay Neural Network (TDNN) [5, 6] is the most currently used architecture. TDNN is the x-vector front-end and consists of a stack of 1-D dilated Convolutional Neural Networks (CNNs). The idea behind the use of TDNNs is to encode a sequence of MFCC into a more discriminative sequence of vectors by capturing long-term feature relations. 2-D CNNs have also shown competitive results for speaker verification. There are Computer Vision architectures such as VGG [7, 10, 11] and ResNet [8, 12, 13] that have been adapted to capture speaker discriminative information from the Mel-Spectrograms. In fact, Resnet34 has shown a better performance than TDNN in the most recent speaker verification challenges [14, 15]. Finally, there are also some other attempts to work directly on the raw signal instead of using hand-crafted features [16, 17, 18].

Given the encoded sequence from the front-end, a pooling layer is adopted to obtain an utterance-level representation. During the last few years, there are several studies addressing different types of pooling strategies [19, 20, 21]. X-vector originally uses statistical pooling [6]. Self attention mechanisms have been used to improve statistical pooling, such as [22]. In works like [23], attention is used to extract better order features statistics. A wide set of pooling layers based on self attention have been proposed improving this vanilla self attention mechanism. In [22] several attentions are applied over the same encoded sequence, producing multiple context vectors. In our previous work [11], the encoded sequence is split into different heads and a different attention model is applied over each head sub-sequence. Non self attention mechanisms have also been proposed like [24], where a mutual attention network is fed with the pair of utterances aimed to compare.

In this paper we present a Double Multi-Head Attention (MHA) pooling layer for speaker verification. The use of this layer is inspired by [25], where Double MHA is presented as a double self attention block which captures feature statistics and makes adaptive feature assignment over images. In this work this mechanism is used as a combination of two self attention pooling layers to create utterance-level speaker embeddings. Given a sequence of encoded representations from a CNN, Self MHA first concatenates the context vector from K head attentions applied over a K sub-embedding sequences. An additional self attention mechanism is then applied over the multi-head context vector. This attention based pooling summarizes the set of head context vectors into a global speaker representation. This representation is pooled through a weighted average of the head context vectors, where the head weights are produced with the self attention mechanism. On the one hand, this approach allows the model to attend to different parts of the sequence, capturing at the same time different subsets of encoded representations. On the other hand, the pooling layer allows to select which head context vectors are the most relevant to produce the global context vector.

This work was supported by the Spanish project PID2019-107579RB-I00 / AEI / 10.13039/501100011033.

In comparison with [25], the second pooling layer operates over the head context vectors produced by a MHA instead of the global descriptors produced by a self multi attention mechanism applied over an image.

2 Proposed Architecture

Our proposed system architecture is illustrated in Figure 1. It utilizes a CNN-based front-end which takes in a set of variable length mel-spectrogram features and outputs a sequence of speaker representations. These speaker representations are further subject to a Double MHA pooling which is the main contribution of this work. The Double MHA layer comprises a Self MHA pooling and an additional Self Attention layer that summarizes the information of each head context vector into a unique speaker embedding. The combination of Self MHA pooling together with this Self Head Attention layer provides us with a deeper self-attention pooling mechanism (Figure 2). The speaker embedding obtained from the pooling layer is sent through a set of FC layers to predict the speaker posteriors. This network architecture is trained with Additive Margin Softmax (AMS) loss [26] as a speaker classifier so as to have a speaker embedding extractor.

2.1 Front-End Feature Extractor

Our feature extractor network is a larger version of the adapted VGG proposed in [11]. This CNN comprises four convolution blocks, each of which contains two concatenated convolutional layers followed by a max pooling with a 2×2 stride. Hence given a spectrogram of N frames, the VGG performs a down-sampling reducing its output into a sequence of $N/16$ representations. The output of the VGG $h \in \mathbb{R}^{M \times N/16 \times D'}$ is a set of M feature maps with $N/16 \times D'$ dimension. These feature maps are concatenated into a unique vector sequence. This reshaped sequence of hidden states can now be defined as $h \in \mathbb{R}^{N/16 \times D}$, where $D = MD'$ corresponds to the hidden state dimension.

2.2 Self Multi-Head Attention Pooling

The sequence of hidden states output from the front-end feature extractor can be expressed as $h = [h_1 h_2 \dots h_N]$ with $h_t \in \mathbb{R}^D$. If we consider a number of K heads for the MHA pooling, now we can define the hidden state as $h_t = [h_{t1} h_{t2} \dots h_{tK}]$ where $h_{tj} \in \mathbb{R}^{D/K}$. Hence each feature vector is split into a set of sub-feature vectors of size D/K . In the same way, we have also a trainable parameter $u = [u_1 u_2 \dots u_K]$ where $u_j \in \mathbb{R}^{D/K}$. A self attention operation is then applied over each head of the encoded sequences. The weights of each head alignment are defined as:

$$w_{tj} = \frac{\exp\left(\frac{h_{tj}^T u_j}{\sqrt{d_h}}\right)}{\sum_{l=1}^K \exp\left(\frac{h_{tl}^T u_l}{\sqrt{d_h}}\right)} \quad (1)$$

where w_{tj} corresponds to the attention weight of the head j on the step t of the sequence and d_h corresponds to hidden state dimension D/K . If each head corresponds to a subspace of the hidden state, the weight sequence of that head can be considered as a probability density function (pdf) from that subspace features over the sequence. We then compute a new pooled representation for each head in the same way than vanilla self attention:

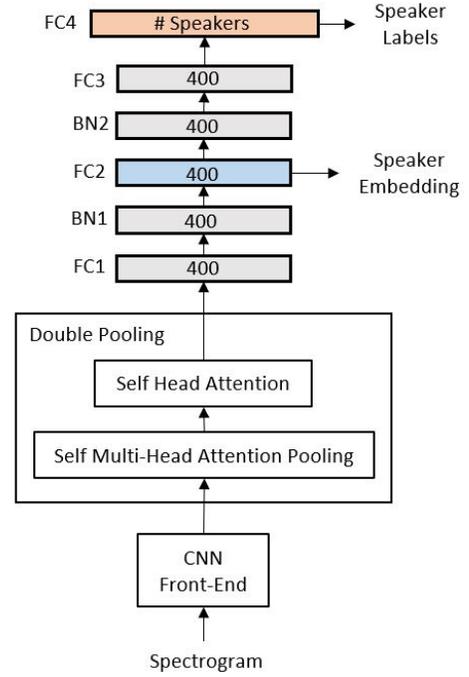


Fig. 1. System Architecture.

$$c_j = \sum_{t=1}^N h_{tj}^T w_{tj} \quad (2)$$

where $c_j \in \mathbb{R}^{D/K}$ corresponds to the utterance level representation from head j . The final utterance level representation is then obtained with the concatenation of the utterance level vectors from all the heads $c = [c_1 c_2 \dots c_k]$. This method allows the network to extract different kinds of information over different regions of the network.

2.3 Double Multi-Head Attention

The main disadvantage of Self MHA pooling is that it assumes uniform head relevance. The output context vector is the concatenation of all head context vectors and it is used as input of the following dense layers. Double MHA does not assume that. Therefore, each utterance context vector is computed as a different linear combination of head context vectors. A summarized vector c is then defined as a weighted average over the set of head context vectors c_i . Self attention is then used to pool the set of head context vectors c_i and obtain an overall context vector c .

$$w'_i = \frac{\exp(c_i^T u')}{\sum_{l=1}^K \exp(c_l^T u')} \quad (3)$$

$$c = \sum_{i=1}^K c_i^T w'_i \quad (4)$$

where w'_i corresponds to the aligned weight of each head and $u' \in \mathbb{R}^{D/K}$ is a trainable parameter. The context vector c is then computed as the weighted average of the context vectors among

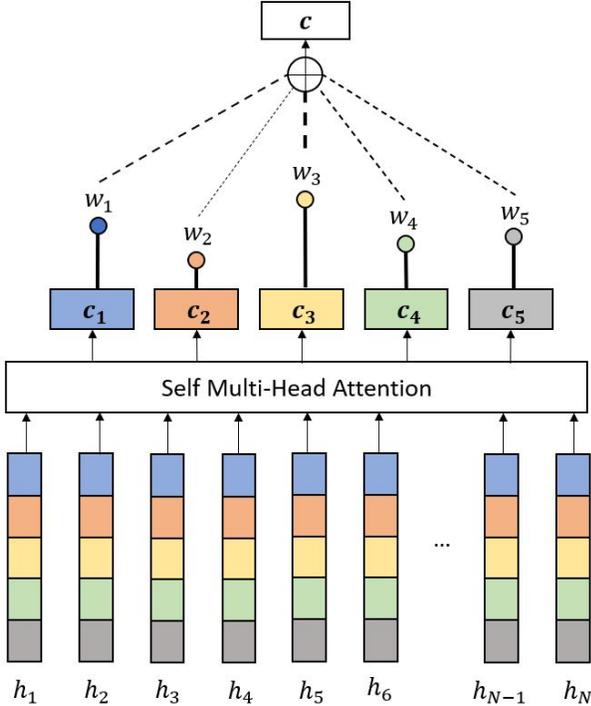


Fig. 2. An example of Double MHA Pooling with 5 heads.

heads. With this method, each utterance context vector is created scaling the information of the most/least relevant heads. Considering the whole pooling layer, Double MHA allows to capture different kind of speaker patterns in different regions of the input, and at the same time allows to weight the relevance of each of these patterns for each utterance.

The number of heads used for this pooling defines both the context vector dimension and how the VGG feature maps are grouped. Considering the number of M channels and K heads, for each head we would create a c_i context vector of $D'M/K$ dimension which contains a subset of M/K feature maps. Therefore, as the number of heads grows larger, it allows Double MHA to consider more subsets of features while decreases the dimension of the final utterance-level context vector. This implies a trade-off between the number of features subsets we can create and how much compressed are these features in the context vector subspace.

2.4 Fully-Connected Layers

The utterance-level speaker vector obtained from the pooling layer is fed into a set of four FC layers (Figure 1). Each of the first two FC layers is followed by a batch normalization [27] and Rectified Linear Unit (ReLU) activations. A dense layer is adopted for the third FC layer and the last FC corresponds to the speaker classification layer. Since AMS is used to train the network, the third layer is set up without activation and batch normalization as proposed in [26]. Once the network is trained, we can extract a speaker embedding from one of the intermediate FC layers. According to [26], we consider the second layer as the speaker embedding instead of the third one. The output of this FC layer then corresponds to the speaker representation that will be used for the speaker verification task.

Table 1. CNN Architecture. In and Out Dim. refers to the input and output feature maps of the layer. Feat Size refers to the dimension of each one of this output feature maps.

Layer	Size	In Dim.	Out Dim.	Stride	Feat Size
conv11	3x3	1	128	1x1	$N \times 80$
conv12	3x3	128	128	1x1	$N \times 80$
mpool1	2x2	-	-	2x2	$N/2 \times 40$
conv21	3x3	128	256	1x1	$N/2 \times 40$
conv22	3x3	256	256	1x1	$N/2 \times 40$
mpool2	2x2	-	-	2x2	$N/4 \times 20$
conv31	3x3	256	512	1x1	$N/4 \times 20$
conv32	3x3	512	512	1x1	$N/4 \times 20$
mpool3	2x2	-	-	2x2	$N/8 \times 10$
conv41	3x3	512	1024	1x1	$N/8 \times 10$
conv42	3x3	1024	1024	1x1	$N/8 \times 10$
mpool4	2x2	-	-	2x2	$N/16 \times 5$
flatten	-	1024	1	-	$N/16 \times 5120$

3 Experimental Setup

The proposed system¹ in this work has been assessed by VoxCeleb dataset [28, 7]. VoxCeleb is a large multimedia database that contains more than one million 16kHz audio utterances for more than 6K celebrities. VoxCeleb has two different versions with several evaluation protocols. For our experiments, VoxCeleb2 development partition with no augmentation have been used to train both baseline and presented approaches. On the other hand, the performance of these systems have been evaluated with Vox1 test, Vox1-E, and Vox1-H conditions. These protocols contain sets of 37,611, 581,480, and 552,536 of Vox1 random pairs, respectively. Vox1 test only uses the test set, Vox1-E uses the whole development + test corpus and Vox1-H is restricted to audio pairs from same nationality and gender speakers.

Two different baselines have been considered to compare with the presented approach. Double MHA pooling have been evaluated against two self attentive based pooling methods: vanilla Self Attention and Self MHA. In order to evaluate them, these mechanisms have replaced the pooling layer of the system (Figure 1) without modifying any other block or parameter from the network. The speaker embeddings used for the verification tests have been extracted from the same FC layer for each of the pooling methods. Cosine distance has been used to compute the scores between pairs of speaker embeddings.

The proposed network has been trained to classify variable-length speaker utterances. As input features we have used 80 dimension log Mel Spectrograms with 25ms length Hamming windows and 10ms window shift. The audio features have been normalized with Cepstral Mean Normalization (CMN). The CNN encoder is then fed with $N \times 80$ spectrograms to obtain a sequence of $N/16 \times 5120$ encoded hidden representations. For training we have used batches of $N = 350$ frames audio chunks but for test the whole utterances have been encoded. The setup of the CNN feature extractor can be found on Table 1. For the pooling layer we have tuned the number of heads for both Self MHA and Double MHA. For the presented CNN setup we have considered 8, 16, and 32 heads, which

¹Models are available at: <https://github.com/miquelindia90/DoubleAttentionSpeakerVerification>

Table 2. Evaluation results on VoxCeleb 1 protocols. Head and global context vectors are referred to as c_i and c .

Approach	Pooling Setup			Vox1 Test		Vox1-E		Vox1-H	
	Heads	c_i dimension	c dimension	EER	DCF	EER	DCF	EER	DCF
Attention	1	5120	5120	3.42	0.0031	3.42	0.0029	4.89	0.0038
MHA	8	640	5120	3.36	0.0029	3.44	0.0029	5.04	0.004
MHA	16	320	5120	3.43	0.0032	3.4	0.0029	4.9	0.004
MHA	32	160	5120	3.64	0.0032	3.68	0.0031	5.35	0.0042
Double MHA	8	640	640	3.27	0.0028	3.23	0.0028	4.69	0.0037
Double MHA	16	320	320	3.19	0.0027	3.22	0.0027	4.67	0.0038
Double MHA	32	160	160	3.23	0.0028	3.18	0.0026	4.61	0.0036

implies a head context vector c_i of 640, 320, and 160, respectively. Models with 64 heads have been discarded due to the instability of their training. The last block of the system consists of four consecutive FC layers. The first three dense layers have 400 dimension. The last FC layer has 5994 dimension, which corresponds to the number of training speaker labels. Batch normalization has been applied only on the first two dense layers as mentioned in subsection 2.4. The network has been trained with AMS loss with $s = 30$ and $m = 0.4$ hyper-parameters. Batch size have been set to 128 and Adam optimizer has been used to train all the models with a $1e-4$ learning rate and a $1e-3$ weight decay. Models have been trained for 100 epochs using a learning rate annealing strategy. After each 15 epochs without validation improvement, learning rate was decayed by a 0.5 factor.

4 Results

The proposed approach has been evaluated against different attention methods in the VoxCeleb text-independent speaker verification task. Performance is evaluated using Equal Error Rate (EER) and Detection Cost Function (DCF) calculated using $C_{FA} = 1$, $C_M = 1$, and $P_T = 0.01$. The results of this task are presented in Table 2. Besides the mentioned metrics, both head c_i and c context vectors dimensions are shown for each presented pooling approach.

Self Attention pooling has shown very similar results compared to Self MHA approaches. In comparison to Self Attention, Self MHA has shown better results in Vox1-test and Vox1-E protocols with 8 heads and 16 heads, respectively. The relative improvement of these approaches compared to Self Attention are 1.75% in terms of EER in Vox1 Test for 8 heads and 0.58% in terms of EER in Vox1-E for the 16 heads model. Otherwise, Self Attention showed the best baseline result in Vox1-H with a 4.89% EER and a 0.0038 DCF. Compared to the best Self MHA approach, Self attention has only shown a relative improvement of 0.58% in terms of EER. This similarity in the results indicates that Self MHA has not led to a noticeable performance improvement compared to vanilla Self Attention pooling. Double MHA have shown better results for all head values compared with both Self Attention and Self MHA approaches. In average, the 32 heads model has outperformed all the baseline systems considering both EER and DCF metrics. For Vox1 Test, Vox1-E and Vox1-H, 32 heads Double MHA have shown a 11.26%, 13.58% and 13.83% of EER relative improvement in comparison to the 32 heads Self MHA model, respectively. This performance increase has also been shown for 8 and 16 heads in average for all the protocols with a 5, 23% and 5, 62% improvement, respectively. Hence Double MHA has provided the best results and it has shown to be more effective than Self MHA for all the head values.

As the results have shown, best performance in Double MHA based models has been achieved with 16 and 32 heads. We have in-

cluded head and global context vector dimensions in Table 2 in order to analyze the relation between the number of heads and the models performance. As it was discussed in subsection 2.3, c_i in Self MHA and both c_i and c dimensions in Double MHA are inversely proportional to the number of heads. Therefore, there is a trade-off between the number of attentions used over the encoded sequence and the amount of speaker information each attention is able to capture. Worst performance with Double MHA was achieved with 8 heads. This setup implies that both c_i and c dimensions are 640. Current state-of-the-art speaker embeddings have a dimension range between 200 and 1500 approximately. This means that there is still some margin to reduce the c dimension and increase the number of attentions used. Increasing the number of heads has led to a better verification performance. The best Double MHA model has 32 heads, whose context vector dimension c is 160. Here the speaker information was encoded into a lower dimension representation and the pooling was allowed to attend to 32 different sub-sets of CNN channels. Models with a larger number of heads have also been considered. However these models could not have been trained due to the narrow c dimension, which led us to unstable training. Therefore Double MHA can be considered as a regularized extension of Self MHA that works as a bottleneck layer. On the other hand, the head number selection is also related to the CNN setup. The results indicate that CNN output feature maps are more efficiently grouped in subsets of $M/K = 32$ channels, which correspond to sub-sequences of 160 dimension embeddings. Considering these sets of 32 context vectors pooled in that layer, these representations are efficiently averaged with Double MHA into unique 160 dimension utterance-level speaker representations.

5 Conclusion

In this paper we have implemented a Double Multi-Head Attention mechanism to obtain speaker embeddings at utterance level by pooling short-term representations. The proposed pooling layer is composed of a Self Multi-Head Attention pooling and a Self Attention mechanism that summarizes the context vectors of each head into a unique speaker vector. This pooling layer has been tested in a neural network based on a CNN. The CNN maps the spectrograms into sequences of speaker vectors. These vectors are the inputs to the proposed pooling layer, whose output activations are then connected to a set of dense layers. The network is trained as a speaker classifier and a bottleneck layer from these fully connected layers is used as speaker embedding. We have evaluated this approach with other pooling methods for the text-independent speaker verification task using the speaker embeddings and applying the cosine distance. The presented approach has outperformed both vanilla Self Attention and Self Multi-Head Attention poolings.

6 References

- [1] Omid Ghahabi, Pooyan Safari, and Javier Hernando, “Deep learning in speaker recognition,” in *Development and Analysis of Deep Learning Architectures*, pp. 145–169. Springer, 2020.
- [2] 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.
- [3] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “Spoken language recognition using x-vectors,” in *Odyssey*, 2018, pp. 105–111.
- [4] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak, “x-vectors meet emotions: A study on dependencies between emotion and speaker recognition,” *arXiv preprint arXiv:2002.05039*, 2020.
- [5] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.
- [6] 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.
- [7] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *INTERSPEECH*, 2018.
- [8] Gautam Bhattacharya, Jahangir Alam, and Patrick Kenny, “Deep speaker recognition: Modular or monolithic?,” in *Proc. Interspeech*, 2019, pp. 1143–1147.
- [9] Pooyan Safari, Miquel India, and Javier Hernando, “Self-attention encoding and pooling for speaker recognition,” *arXiv preprint arXiv:2008.01077*, 2020.
- [10] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.
- [11] Miquel India, Pooyan Safari, and Javier Hernando, “Self Multi-Head Attention for Speaker Recognition,” .
- [12] Jianfeng Zhou, Tao Jiang, Zheng Li, Lin Li, and Qingyang Hong, “Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function,” *Proc. Interspeech 2019*, pp. 2883–2887, 2019.
- [13] Amirhossein Hajavi and Ali Etemad, “A deep neural network for short-segment speaker recognition,” *arXiv preprint arXiv:1907.10420*, 2019.
- [14] Joon Son Chung, Arsha Nagrani, Ernesto Coto, Weidi Xie, Mitchell McLaren, Douglas A Reynolds, and Andrew Zisserman, “Voxsrc 2019: The first voxceleb speaker recognition challenge,” *arXiv preprint arXiv:1912.02522*, 2019.
- [15] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, “But system description to voxceleb speaker recognition challenge 2019,” *arXiv preprint arXiv:1910.12592*, 2019.
- [16] Mirco Ravanelli and Yoshua Bengio, “Speaker recognition from raw waveform with sincnet,” *arXiv preprint arXiv:1808.00158*, 2018.
- [17] Jee-Weon Jung, Hee-Soo Heo, IL-Ho Yang, Hye-Jin Shim, and Ha-Jin Yu, “Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification,” *extraction*, vol. 8, no. 12, pp. 23–24, 2018.
- [18] 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.
- [19] Weicheng Cai, Jinkun Chen, and Ming Li, “Exploring the encoding layer and loss function in end-to-end speaker and language recognition system,” in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 74–81.
- [20] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Utterance-level aggregation for speaker recognition in the wild,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.
- [21] Youngmoon Jung, Younggwan Kim, Hyungjun Lim, Yeunju Choi, and Hoirin Kim, “Spatial pyramid encoding with convex length normalization for text-independent speaker verification,” *arXiv preprint arXiv:1906.08333*, 2019.
- [22] Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey, “Self-attentive speaker embeddings for text-independent speaker verification,” in *Interspeech*, 2018, pp. 3573–3577.
- [23] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, “Attentive statistics pooling for deep speaker embedding,” *arXiv preprint arXiv:1803.10963*, 2018.
- [24] Jingyu Li and Tan Lee, “Text-independent speaker verification with dual attention network,” *arXiv preprint arXiv:2009.05485*, 2020.
- [25] Yungpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng, “A²-nets: Double attention networks,” in *Advances in Neural Information Processing Systems*, 2018, pp. 352–361.
- [26] Yi Liu, Liang He, and Jia Liu, “Large margin softmax loss for speaker verification,” *arXiv preprint arXiv:1904.03479*, 2019.
- [27] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.
- [28] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” in *INTERSPEECH*, 2017.