Speaker Verification Experiments for Adults and Children using a shared embedding spaces

Tuomas Kaseva, Hemant Kathania, Aku Rouhe, Mikko Kurimo
Department of Signal Processing and Acoustics, Aalto University, Finland
(firstname.lastname)@aalto.fi

Abstract

In this work, we present our efforts towards developing a robust speaker verification system for children when the data is limited. We propose a novel deep learning -based speaker verification system that combines long-short term memory cells with NetVLAD and additive margin softmax loss. First we investigated these methods on a large corpus of adult data and then applied the best configuration for child speaker verification. For children, the system trained on a large corpus of adult speakers performed worse than a system trained on a much smaller corpus of children's speech. This is due to the acoustic mismatch between training and testing data. To capture more acoustic variability we trained a shared system with mixed data from adults and children. The shared system yields the best EER for children with no degradation for adults. Thus, the single system trained with mixed data is applicable for speaker verification for both adults and children.

Index Terms: additive margin softmax loss, NetVLAD aggregation, recurrent neural network, speaker verification for children.

1 Introduction

The use of speaker verification (SV) technology for children has many beneficial application areas, such as child security and protection, entertainment, games and education. For example, in an interactive class the teacher could identify each child, by continuing a previous lecture and adapt its content with the child's speech, and log the child's responses without a conventional login process (Safavi et al., 2018, 2012).

The acoustic and linguistic characteristic of children's speech differ from adults' speech (Lee et al., 1999). The main differences are in pitch, speaking rate and formant frequencies (Kumar Kathania et al., 2020; Shahnawazuddin et al., 2019). These acoustic differences together with the lack of training data make SV more challenging. Little work has been reported in this area. In (Shahnawazuddin et al., 2020) in-Domain and out-of-Domain data augmentation are used to improve a child SV system in a limited data scenario. In (Safavi et al., 2012) vocal tract information is used for children's SV. Explanation for degraded recognizer scores through acoustic changes resulting from voice disguise is presented in (González Hautamäki et al., 2019).

In this work, we explore how recent advances in (adult) SV could aid in child SV, as well. In particular, we combine adult and child SV into a single task, by using a shared embedding space for adult and child speakers. This allows us to leverage the large resources available for adult speakers for the low-resource child speaker verification task. In applications where both adult and child speakers can be expected, it is also natural to use a single shared system for both groups; we find that a shared system can be used which benefits child SV without degrading adult SV performance.

Contributions. We construct a neural SV system, which leverages recent advancements in the field. In particular, we find improvements from using the additive-margin softmax loss and the NetVLAD time aggregation methods. In contrast to most recent literature, which uses convolutional neural layers, we apply recurrent layers, motivated by success in speaker diarization(Kaseva et al., 2019). We compare our results to recent high-performing systems of similar complexity. Though we do not outperform the top results, the comparison validates our approach. We then apply the proposed SV system to adult and child SV

and find that using a shared embedding for both adult and children improves child SV drastically without affecting adult SV performance.

2 Related speaker verification work

In recent years, deep learning motivated approaches have shown significant progress in SV. We consider three main reasons for their success. Firstly, larger and more realistic speakers-in-the-wild speaker recognition datasets have become available to the public (Nagrani et al., 2017; Chung et al., 2018; McLaren et al., 2016). Secondly, the loss functions used in the training of neural networks have advanced. In general, the main objective of the neural networks designed for SV is to transform a given recording into a speaker embedding which embodies the speaker characteristics of the recording (Snyder et al., 2018, 2019; Bredin, 2017; Li et al., 2017). In the most current methods, the embeddings are learned in a speaker identification process, were original softmax loss is modified by adding a margin to the class decision boundaries (Liu et al., 2019; Xie et al., 2019; Xiang et al., 2019) . This allows efficient training and reduces the intra-class variance of the created embeddings (Wang et al., 2018a; Deng et al., 2019; Liu et al., 2017). Finally, the neural network architectures have developed. One of the most prominent discoveries has been x-vectors, speaker embeddings which are extracted from an architecture based on time-delay neural networks (TDNNs) (Snyder et al., 2018; Liu et al., 2019; Xiang et al., 2019). X-vectors have been shown to outperform i-vectors, which have enjoyed a state-of-the-art status in SV for a long time (Dehak et al., 2010). In some cases, i-vectors have also been inferior to the SV systems which utilize convolutional neural networks (CNNs) (Chung et al., 2018; Ravanelli and Bengio, 2018). Furthermore, novel aggregation methods for neural networks have been proposed. Whereas average pooling has been used extensively before, the most recent approaches include statistics pooling, attentive statistics pooling and NetVLAD (vector of locally aggregated descriptors) (Okabe et al., 2018; Arandjelovic et al., 2016; Xie et al., 2019).

In addition, recurrent neural networks (RNNs) with long-short term memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) have been experimented with (Wan et al., 2018; Bredin, 2017; Heigold et al., 2016). Most importantly, they have shown success in a related task, online speaker diarization (Wang et al., 2018c; Zhang et al., 2019; Wisniewksi et al., 2017). In this task, LSTMs have been been able to create compact speaker embeddings from very short segments.

Our approach has some similarities with Wan et al. (Wan et al., 2018). As in their work, we use LSTMs in sliding windows. However, unlike them, we do not apply the generalized end-to-end loss for neural network training. Instead, we use the AM-softmax loss (Wang et al., 2018a). Furthermore, unlike them, we combine LSTMs with NetVLAD. Although NetVLAD layer has been previously used for SV (Xie et al., 2019), in that study, the layer was connected to a CNN. NetVLAD has been originally designed for aggregation of CNNs (Arandjelovic et al., 2016) and to the best of our knowledge, we are the first to use it with LSTMs in any application.

3 Proposed methods

In this section, we detail our SV system which consists of three stages: splitting, embedding and averaging, as illustrated in Fig. 1.

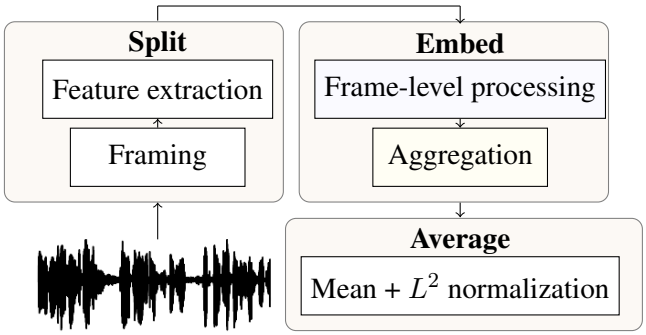

Figure 1: Schematic of our speaker embedding extraction approach.

Split. First, the audio input is split into overlapping windows with short, roughly 2 seconds or less, duration. Time-varying features are then extracted from each frame, resulting in a set of feature sequences $\mathbf{x}$. The sequences consist of 30 Mel-Frequency Cepstral Coefficients (MFCC) which are extracted every 10ms with 25ms frame length. Every $\mathbf{x}$ is normalized with zero mean and unit variance.

Embed. In the next step, each $\mathbf{x}$ is transformed into a speaker embedding. This can be further divided into two distinct steps: frame-level processing and aggregation.

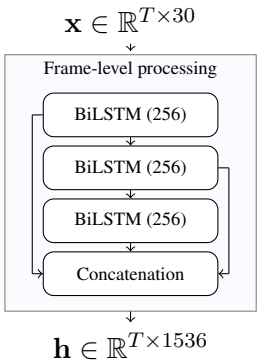

Figure 2: Frame-level processing. The numbers refer to the number of hidden units in each layer.

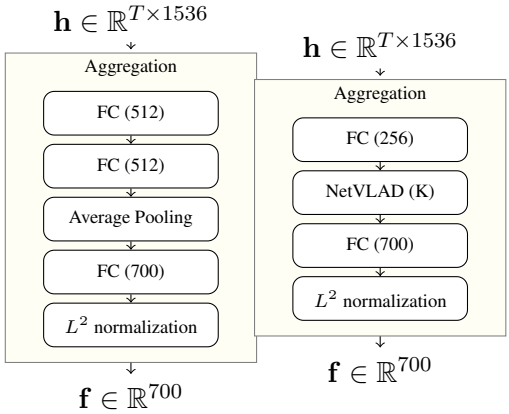

Figure 3: Two different aggregation approaches: average pooling on the left and NetVLAD on the right. FC refers to a fully connected layer and the numbers to the output dimensionality.

In frame-level processing, each $\mathbf{x}$ is transformed into higher level frame-features $\mathbf{h}$. In our approach, $\mathbf{x}$ is fed to a cascade of three bidirectional LSTM layers with skip connections. Each layer outputs the hidden states of both the forward and backward LSTMs. These outputs are concatenated resulting in $\mathbf{h}$ as illustrated in Figure 2. The structure of the cascade adheres to (Wisniewksi et al., 2017). A more common choice for frame-level processing blocks is to use convolutional layers.

In aggregation, the higher level features $\mathbf{h}$ are compressed into a speaker embedding $\mathbf{f}$. We compare two aggregation approaches: average pooling and NetVLAD. The aggregation components are illustrated in Fig. 3. Note that the aggregation component with average pooling has a slightly different configuration than its NetVLAD motivated counterpart. This choice was based on balancing the number of parameters in both neural networks.

We force the embeddings to be L^2 normalized in both components. As a result, cosine distance is the most natural distance metric between different embeddings. A rectified linear unit activation is used in all of the fully connected (FC) layers. We also apply batch normalization (Ioffe and Szegedy, 2015) after each layer except L^2 normalization layers. This means that the last two layers of both components perform normalization. Although this might seem strange, we discovered it to be beneficial in the preliminary experiments.

The operation of the NetVLAD layer can be summarized as follows. Let us denote the output of the preceding FC layer as $\mathbf{v} \in \mathbb{R}^{T\times 256}$. First, $\mathbf{v}$ is transformed into $\mathbf{V} \in \mathbb{R}^{K\times 256}$ according to a formula (Arandjelovic et al., 2016)

$$\mathbf{V}(k,d) = \sum_{t=1}^{T} \frac{e^{\mathbf{w}_k^T \mathbf{v}_t + b_k}}{\sum_{k'=1}^{K} e^{\mathbf{w}_{k'}^T \mathbf{v}_t + b_{k'}}} (\mathbf{v}_{td} - \mathbf{c}_{kd}), \tag{1}$$

where $\mathbf{c} \in \mathbb{R}^{K\times 256}$, $\mathbf{w} \in \mathbb{R}^{K\times 256}$ and $\mathbf{b} \in \mathbb{R}^{K}$ are learnable parameters. In this formulation, $\mathbf{c}$ can be interpreted as a set of K cluster centers which characterize the distribution of $\mathbf{v}$ (Xie et al., 2019). More specifically, $\mathbf{V}$ consists of first order statistics of residuals $\mathbf{v}_d - \mathbf{c}_k$ in which each element is weighted based on $\mathbf{v}$ and the cluster index k. The number of clusters K is given as an input to the layer. After calculation of the residuals, each row of $\mathbf{V}$ is first L^2 normalized and then concatenated resulting in $\mathbf{V_f} \in \mathbb{R}^{256*K}$. In the literature, additional L^2 normalization operation has been applied after flattening (Xie et al., 2019; Arandjelovic et al., 2016). However, we use batch normalization instead. We found this normalization to perform generally better in the preliminary experiments. The use of NetVLAD in this study is motivated by its recent success in SV when combined with CNNs (Xie et al., 2019). Here, we show that NetVLAD is beneficial also with LSTMs.

Average. In the final stage, we compute a single embedding $\mathbf{f}_c \in \mathbb{R}^{700}$ for the recording by averaging the created speaker embeddings and L^2 normalizing the average. When considering $\mathbf{f}_{c1}$ and $\mathbf{f}_{c2}$ extracted from two different recordings, our system performs SV by computing cosine distance between the embeddings and by thresholding the obtained value. Another popular method for comparing the embeddings is Probabilistic Discriminant Analysis (PLDA) (Ioffe, 2006; Snyder et al.,

2018). PLDA could result in performance improvements (Liu et al., 2019), but also increase the complexity of our system, and we do not apply it in this work.

4 Experiments

Data. We use two training sets for adult speakers which are both generated from Voxceleb2 (Chung et al., 2018). In the first, abbreviated as $VC2$, all recordings in Voxceleb2 are windowed into 2 second samples with 1 second overlap. The reason for this choice is the training objective of our neural networks that is to identify a speaker from a given training set based on a 2 second segment of speech. The duration was not selected arbitrarily: we experimented also with setting it to 1 and 2.5 seconds. The former was too short for neural networks to learn speaker characteristics properly and the latter did not generally improve the performance of the networks. $VC2$ consists of roughly 6.83 million training samples from 5994 speakers.

The second set, $VC2_C$, is otherwise the same as $VC2$ but excludes a portion of the samples based on a heuristic cleaning algorithm. The motivation for this algorithm came from our listening tests which confirmed that Voxceleb2 included wrongly labeled speaker identities in some cases. The exclusions removed approximately 46k samples from $VC2$ but retained the number of speakers, 5994. Given samples S_i belonging to i-th speaker in $VC2$, the cleaning algorithm operates in four steps:

1. Create a speaker embedding $\mathbf{f}$ for each sample in S_i.

2. Cluster the embeddings with spherical K-means setting $K = 2$ into groups G_1 and G_2.

3. Calculate the average of silhouette coefficients ϕ of the clustering result. Further details of these coefficients are given in (Rousseeuw, 1987).

4. If $|G_1| > 0.6|G_1 \cup G_2|$ and $\phi > 0.3$, exclude all samples belonging to G_2 from the training set. Here, $|G_i|$ refers to a number of elements in group G_i.

In summary, the algorithm investigates whether the recordings initially assigned to a single speaker might contain also another speaker. The algorithm removes samples from S_i only if the speech material portions of the clusters are not balanced and if the clustering result has a high reliability. This reliability is measured using silhouette coefficients. Speaker embedding extraction was performed using an initial neural network which has the same average pooling based architecture as described in the previous section, but was trained only with 4000 speakers from $VC2$.

We evaluate our models also using the cleaned versions of Voxceleb1 verification test sets, Voxceleb1-test (VC_t), Voxceleb1-H (VC_H) and Voxceleb1-E (VC_E) (Chung et al., 2018). The recordings in these sets are framed to 2 second duration segments with 1.5 seconds overlap. The overlap duration was determined in the preliminary experiments.

We construct also our own verification set from the development set of Voxceleb1. This set is used for model evaluation during training. The set consists of speech segments with a fixed 2 seconds duration, and which each are extracted from a unique session and speaker. The number of extracted segments is close to 20k and they belong to 1211 speakers. We form close to 150k segment pairs where half of the pairs correspond to the same speaker and the other half to different speakers. We name this verification set as VC_{2sec}. The set is disjoint in speakers with VC_t but not with VC_H and VC_E. However, we consider this evaluation set to be valid since the pair compositions and segment durations of VC_{2sec} differ significantly from VC_H and VC_E.

For child speech experiments we used CSLU kids (Khaldoun Shobaki, 2007) database for training. It has 1110 speakers of English language with age range from 6 to 16 years and sampling rate 16 kHz. For testing the system we used PF-STAR (Batliner et al., 2005) and the Finnish SpeechDat (Rosti et al., 1998) datasets. PFSTAR has 134 speakers of English with age range from 4 to 14 years, originally sampled at 22,050 Hz. The down-sampling at 16 kHz was performed for consistency with the model. SpeechDat has 354 speakers of Finnish with age range from 6 to 14 years, originally data sampled at 8 kHz. The up-sampling at 16 kHz was performed for consistency. For children's experiments we used the same speaker embedding method as adults.

Training. In training, the output of the aggregation component is connected to a fully connected

layer which is used for a speaker identification task. Training has two stages: warm-up with the softmax loss and fine-tuning with the AM-softmax loss (Wang et al., 2018a). In the warm-up, the neural network is trained for 5 epochs, using Adam optimizer with 0.01 learning rate. Batch size is chosen as 512. We generally observed that the performance of the neural networks on the VC_{2sec} would not improve after the fifth epoch when using the softmax loss.

In the fine-tuning, the softmax loss for i-th training sample is reformulated as

$$L_i = log\frac{e^{s(\mathbf{W}_{y_i}^T\mathbf{f}-m)}}{e^{s(\mathbf{W}_{y_i}^T\mathbf{f}-m)} + \sum_{j=1,j\neq y_i}^{5994} e^{s\mathbf{W}_j^T\mathbf{f}}}, \quad (2)$$

where y_i is the label of i-th training sample, $\mathbf{W} \in \mathbb{R}^{700\times5994}$ a learnable weight matrix with all rows L^2 normalized and s and m a given scale and margin. Equation 2 is known as the AM-softmax loss (Wang et al., 2018a). We set $m = 0.15$ and $s = 0.25$ based on our preliminary experiments. $\mathbf{W}$ is initialized with the weights of the best neural network configuration found in the warm-up.

The main point of using the AM-softmax loss is to decrease intra-class variance, which is generally difficult with the softmax loss (Wang et al., 2018a; Deng et al., 2019; Liu et al., 2017). In other words, the higher the margin m is set, the more closer, in terms of cosine distance, the speaker embeddings belonging to the same class are forced. The cosine distance metric arises from the L^2 normalizations of both $\mathbf{f}$ and the rows of $\mathbf{W}$. The scale of s is generally set to a some high value to ensure convergence (Wang et al., 2018b). In recent years, the AM-softmax loss and other similar methods (Liu et al., 2017; Deng et al., 2019) have emerged as state-of-the-art approaches in speaker verification (Xie et al., 2019; Liu et al., 2019; Xiang et al., 2019; Li et al., 2018).

The fine tuning is continued for 10 epochs with otherwise the same setting as in warm-up. We monitor the progress of the training by first computing cosine distances between the embeddings of each pair in VC_{2sec} and then calculating equal error rate (EER) on these distances after each epoch. EER is a standard error metric in speaker verification (Snyder et al., 2018; Chung et al., 2018; Xie et al., 2019). Although the VC_{2sec} contains over 150k pairs, the evaluation on this set is efficient during the training since it consists of short, equal length segments which can be embedded rapidly. We save the weights of the neural network after each epoch, and choose the configuration with the best EER value as our final model.

5 Results

First in section 5.1 we validate our SV approach on adult speech. Then in section 5.2 we apply the system with children.

5.1 Validation experiments with adults

In this section, we first investigate the effect of the cleaning algorithm, aggregation and the AM-softmax loss. Finally, we present a results comparison. We use EER as an evaluation metric in all experiments.

Table 1: Effect of training set cleaning (EER %). $K = 30$.

Aggregation	Training set	VC_t	VC_E	VC_H	VC_{2sec}
NetVLAD	$VC2$	2.49	2.47	4.53	**6.65**
NetVLAD	$VC2_C$	**2.18**	**2.45**	**4.45**	6.66

Effect of dataset cleaning. In Table 1, we show that small improvements can be achieved by removing some training data with the cleaning algorithm. This proves that the algorithm is reasonable and also encourages discussion whether some cleaning operation is needed for Voxceleb2. However, the improvements in VC_E and VC_H are minor and with VC_{2sec}, the cleaning has not been beneficial.

Table 2: Effect of K and aggregation (EER %). The training set is $VC2_C$.

Aggregation	K	VC_t	VC_E	VC_H	VC_{2sec}
Average pooling	-	2.46	2.45	4.42	7.05
NetVLAD	8	2.41	2.40	**4.35**	6.92
NetVLAD	14	2.32	**2.37**	4.36	6.68
NetVLAD	30	**2.18**	2.45	4.45	**6.66**

Effect of aggregation approach. Table 2 investigates the performance of the two aggregation approaches and the choice of K. The results show that NetVLAD is the better approach. This is particularly clear with VC_{2sec}. However, the best scores with different test sets are all obtained with different K values. This result highlights the importance of using multiple different test sets for model evaluation. Nevertheless, we can decide on the best model based on the average over all EER scores. In this case, the NetVLAD-based aggregation with $K = 14$ has the best performance.

Table 3: Effect of loss function (EER %). $K = 30$ and the training set is $VC2_C$.

Aggregation	Loss	VC_t	VC_E	VC_H	VC_{2sec}
NetVLAD	Softmax	3.25	3.30	5.90	8.40
NetVLAD	AM-softmax	**2.18**	**2.45**	**4.45**	**6.66**

Effect of loss function. Table 3 illustrates that the AM-softmax loss brings significant improvements over the softmax loss. Similar results were obtained with the average pooling aggregation. However, we want to emphasize the results with the NetVLAD aggregation since in (Xie et al., 2019), the use of NetVLAD with the AM-softmax loss has not resulted in notable performance improvements. Here, we demonstrate that the two can be combined successfully. The results with different K values were essentially the same.

Table 4: Results comparison (EER %).

System	Scoring	VC_t	VC_E	VC_H
Xie *et al.* (Xie et al., 2019)	Cosine	3.22	3.13	5.06
Xiang *et al.* (Xiang et al., 2019)	PLDA	2.69	2.76	4.73
Ours	Cosine	2.32	2.37	4.36
Zhou *et al.* (Zhou et al., 2019)	*Unknown*	2.23	2.18	3.61
Zeinali *et al.* (Zeinali et al., 2019)	Cosine	1.42	1.35	2.48

Results comparison. In Table 4, we compare our system to other high-performing speaker verification systems. The comparison of our system with the first, x-vector based (Xiang et al., 2019) system and the second, CNN-based (Xie et al., 2019) system is straight-forward since all the systems are trained with the same dataset, Voxceleb2, and because the number of parameters are close to each other: 4.2 million in (Xiang et al., 2019), 7.7 million in (Xie et al., 2019) and 6.7 million in our system. Zhou *et al.* (Zhou et al., 2019) report better results than ours, but they use data augmentation, and do not report the number of parameters used. The current state-of-the-art (single-system) results by Zeinali *et al.* (Zeinali et al., 2019) in the VoxCeleb Speaker Recognition Challenge 2019 leverage data augmentation and more parameters. Our results do not outperform the best published results, but the results still validate our approach, as our results outperform the strong results from (Xie et al., 2019) and (Xiang et al., 2019), which use similar parameter and data constraints.

5.2 Evaluation experiments with children

In the previous section, we presented the effect of dataset cleaning, aggregation approach, and loss function on adult speakers. In this section, we took the best combination of all these for child speech experiments. Details of databases used for the experiments with children is given in section 4.

Table 5: Results on child speakers (EER %).

Training data	PF-STAR	Speechdat	VC_E	VC_H
Adults' $VC2_C$	2.58	10.68	2.37	4.36
Children's CSLU	2.05	10.08	–	
Adults' + Children's	**1.12**	**8.82**	2.34	4.39

Table 5 illustrates the performance on child speakers in English and Finnish languages when directly using the adults' model. We also trained a similar model on child speech and report the results in the same table. Based on these results it can be noted that when the system is only trained with adults the performance is lower compared to the system trained only with children even though the children have less training data. To capture more acoustic variability of speakers we trained a shared system with mixed data of both adults and children and tested it with English and Finnish children. The last result of table 5 illustrates that the shared model outperforms both the adult-only and the child-only models for both English and Finnish languages. When compared to a recent paper (Shahnawazuddin et al., 2020) for PF-STAR verification set, our system gives a 50 % relative improvement. Furthermore, when we run the adult test sets VC_E and VC_H again using the final system trained with shared system with mixed of adults' and children's data, we found out that the performance remains the same. This means that we can now use the same model for recognizing both adults and children.

6 Conclusion

We have presented a speaker verification system based on a shared neural embedding space for adults and children. The neural network consists of a cascade of LSTM layers and a NetVLAD aggregation layer, and uses the AM-softmax loss in training. We have demonstrated that the system achieves promising results with adults and children. Because the child data is limited, we trained a shared system with mixed adult and child data to capture more acoustic variability. The shared sys-

tem gives a 54% and 43% relative improvement for children compared to the separate children's and adult systems. For adults, the shared system gives the same performance as compared to the adult system. Finally, we can conclude that this shared system can be now used for both children and adults.

7 Acknowledgment

This work was supported by the Academy of Finland (grant 329267) and EU's Horizon 2020 research and innovation programme via the project MeMAD (GA 780069).

References

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307.

A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong. 2005. The PF_STAR children's speech corpus. In *Proc. INTERSPEECH*, pages 2761–2764.

Hervé Bredin. 2017. Tristounet: triplet loss for speaker turn embedding. In *Proc. ICASSP*, pages 5430–5434. IEEE.

J. S. Chung, A. Nagrani, and A. Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Proc. INTERSPEECH*.

Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.

Rosa González Hautamäki, Ville Hautamäki, and Tomi Kinnunen. 2019. On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise. *The Journal of the Acoustical Society of America*, 146(1):693–704.

Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. 2016. End-to-end text-dependent speaker verification. In *Proc. ICASSP*, pages 5115–5119. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sergey Ioffe. 2006. Probabilistic linear discriminant analysis. In *European Conference on Computer Vision*, pages 531–542. Springer.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456.

T. Kaseva, A. Rouhe, and M. Kurimo. 2019. Spherediar: An effective speaker diarization system for meeting data. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 373–380.

Ronald Allan Cole Khaldoun Shobaki, John-Paul Hosom. 2007. CSLU: Kids' Speech Version 1.1 LDC2007S18. Web Download. Philadelphia. In *Linguistic Data Consortium*.

H. Kumar Kathania, Sudarsana Reddy Kadiri, P. Alku, and M. Kurimo. 2020. Study of formant modification for children asr. In *Proc. ICASSP*, pages 7429–7433.

Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468.

Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. 2017. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*.

Yutian Li, Feng Gao, Zhijian Ou, and Jiasong Sun. 2018. Angular softmax loss for end-to-end speaker verification. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 190–194. IEEE.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220.

Yi Liu, Liang He, and Jia Liu. 2019. Large margin softmax loss for speaker verification. *arXiv preprint arXiv:1904.03479*.

Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. 2016. The speakers in the wild (SITW) speaker recognition database. In *Proc. INTERSPEECH*, pages 818–822.

A. Nagrani, J. S. Chung, and A. Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. In *Proc. INTERSPEECH*.

Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. 2018. Attentive statistics pooling for deep speaker embedding. *arXiv preprint arXiv:1803.10963*.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with sincnet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028. IEEE.

Antti Rosti, Anssi Ramo, Teemu Saarelainen, and Jari Yli-Hietanen. 1998. Speechdat finnish database for the fixed telephone network. *Tech. Rep., Tampere University of Technology*.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Saeid Safavi, Maryam Najafian, Abualsoud Hanani, Martin Russell, Peter Jancovic, and Michael Carey. 2012. Speaker recognition for children's speech. In *Proc. INTERSPEECH*, volume 3.

Saeid Safavi, Martin Russell, and Peter Jančovič. 2018. Automatic speaker, age-group and gender identification from children's speech. *Computer Speech Language*, 50:141 – 156.

S. Shahnawazuddin, Nagaraj Adiga, B Tarun Sai, Waquar Ahmad, and Hemant K. Kathania. 2019. Developing speaker independent asr system using limited data through prosody modification based on fuzzy classification of spectral bins. *Digital Signal Processing*, 93:34 – 42.

S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar. 2020. In-domain and out-of-domain data augmentation to improve children's speaker verification system in limited data scenario. In *Proc. ICASSP*, pages 7554–7558.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Alan McCree, Daniel Povey, and Sanjeev Khudanpur. 2019. Speaker recognition for multi-speaker conversations using x-vectors. In *Proc. ICASSP*, pages 5796–5800. IEEE.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *Proc. ICASSP*, pages 5329–5333. IEEE.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *Proc. ICASSP*, pages 4879–4883. IEEE.

Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. 2018a. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018b. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274.

Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018c. Speaker diarization with lstm. In *Proc. ICASSP*, pages 5239–5243. IEEE.

Guillaume Wisniewksi, Hervé Bredin, Grégory Gelly, and Claude Barras. 2017. Combining speaker turn embedding and incremental structure prediction for low-latency speaker diarization. In *Proc. INTERSPEECH*.

Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu. 2019. Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. *arXiv preprint arXiv:1906.07317*.

Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2019. Utterance-level aggregation for speaker recognition in the wild. In *Proc. ICASSP*, pages 5791–5795. IEEE.

Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot. 2019. But system description to voxceleb speaker recognition challenge 2019.

Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. 2019. Fully supervised speaker diarization. In *Proc. ICASSP*, pages 6301–6305. IEEE.

Tianyan Zhou, yong zhao, Jinyu Li, Yifan Gong, and Jian Wu. 2019. Cnn with phonetic attention for text-independent speaker verification. In *Automatic Speech Recognition and Understanding Workshop*. IEEE.