MULTI-SCALE SPEAKER DIARIZATION WITH NEURAL AFFINITY SCORE FUSION

Tae Jin Park, Manoj Kumar and Shrikanth Narayanan

University of Southern California

ABSTRACT

Predicting the speaker's identity of short speech segments in human dialogue has been considered one of the most challenging problems in speech signal processing. Speaker representations of short speech segments tend to be unreliable, resulting in poor fidelity of speaker representations in tasks requiring speaker recognition. In this paper, we propose an unconventional method that tackles the trade-off between temporal resolution and the quality of the speaker representations. To find a set of weights that balance the scores from multiple temporal scales of segments, a neural affinity score fusion model is presented. Using the CALLHOME dataset, we show that our proposed multi-scale segmentation and integration approach can achieve a state-of-the-art diarization performance.

Index Terms— Speaker Diarization, Uniform Segmentation, Multi-scale, Score Fusion

1. INTRODUCTION

Speaker diarization aims to cluster and label the regions in the temporal domain in terms of speaker identity. In general, the speaker diarization pipeline consists of speech activity detection (SAD), segmentation, speaker representation extraction, and clustering. The segmentation process largely determines the accuracy of the final speaker label because the segmentation determines the unit of diarization output that cannot be altered during the clustering process. In terms of segmentation, a speaker representation faces an inevitable trade-off between the temporal accuracy and speaker representation quality. It has been shown in many previous studies that the speaker representation accuracy improves as the segment length increases [1]. However, specifically in the context of speaker diarization, a longer segmentation means a lower resolution in the temporal domain because a segment is the unit of the process that determines the speaker identity.

In the early days of speaker diarization, the clustering process was based on Bayesian information criterion (BIC) [2], which employs Mel-frequency cepstral coefficients (MFCCs) as a representation for speaker traits. With BIC-based clustering and MFCCs, speech segmentation techniques [3] with a variable segmentation length have been employed because the benefit of having a proper segment length for input speech outweighs the performance degradation from variable segment lengths. This trend has changed with the increase in newer speaker representation techniques, such as i-vector [4, 5] and x-vector [6, 7], where fixing the length of the segments improves the speaker representation quality and reduces additional variability. For this reason, many previous studies have made a point of compromise at 1.0 [8] to 1.5 s [7, 9] depending on the domains they target. However, a fixed segment length has inherent limitations in terms of the temporal resolution because the clustering output can never be finer than the predetermined segment duration.

Therefore, we propose a scheme that addresses the problem arising from such a trade-off and applies a new segmentation approach. The proposed method employs a multi-scale diarization solution where affinity scores from multiple scales of segmentation are fused using a neural score fusion system. The graph in Fig. 1 shows the trade-off between segment length and fidelity of speaker representations from two segments. Our goal is for our system to be located on the graph above the trade-off curve with a higher temporal resolution while at the same time achieving a superior accuracy of the affinity measure.

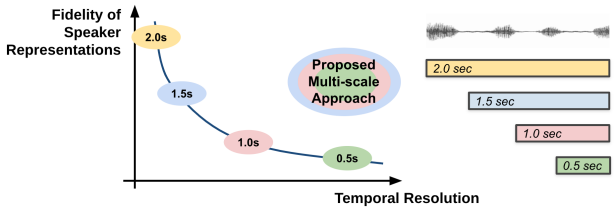

Fig. 1. Trade-off curve between fidelity of speaker representations and temporal resolution.

There have been few studies related to the problem discussed herein. In terms of speaker embedding extraction, few studies have employed a multi-scale concept for speaker embedding [10, 11] targeting short utterance lengths. These studies apply multi-scale aggregation [10] or multilevel pooling [11] in the feature level in the neural network models. Because our proposed neural network model does not generate speaker embeddings, feature-level multi-scale approaches are far from our focus.

By contrast, there are a few studies in which diarization systems aggregate the output of multiple modules. In [12], the authors employed a majority voting scheme on multiple segmentation streams. In [13], the authors introduced a cluster matching procedure that can integrate multiple diarization systems. In addition, in [14], a diarization output voting error reduction (DOVER) was presented for improving the diarization of a meeting speech. These previous studies deal with either a feature-level multi-scale concept of a neural network [10, 11] or a diarization system integration [12–14], whereas our proposed method focuses on the score fusion of multi-scale speech segments.

Our proposed approach[1] has the following novelties. First, unlike conventional varying-length speech segmentation or single-scale segmentation modules, our system employs multiple discrete segment lengths and proposes a method to integrate the given scales. Second, the proposed method can attentively weigh the affinity from multiple scales depending on the domain and characteristics of the given speech signal. This distinguishes our work from approaches that require fusion parameters to be manually tuned on a development set. [15, 16]. In addition to these novelties, our proposed multi-scale approach outperforms a single-scale diarization system and achieves a state-of-the-art performance on the CALLHOME diarization dataset.

The remainder of this paper is structured as follows. In Section

[1] Source code: github.com/tango4j/Multiscale-Speaker-Diarization

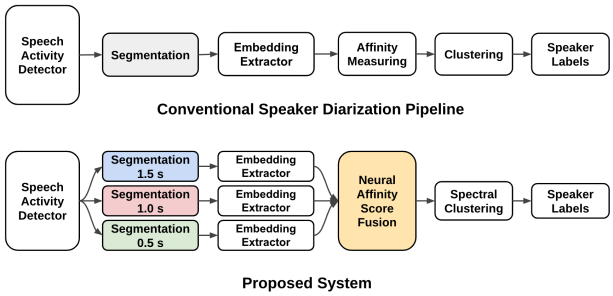

Fig. 2. Example of multi-scale segmentation scheme.

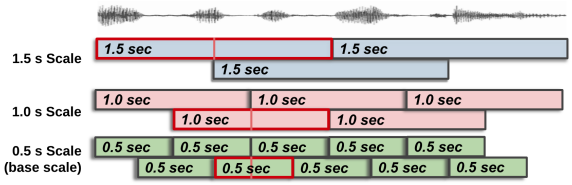

Fig. 3. Example of multi-scale segmentation and mapping scheme.

2, we introduce the segmentation scheme. In Section 3, we introduce the proposed network architecture and how we train the proposed neural network. In Section 4, we show the experimental results on various datasets and evaluation settings.

2. MULTI-SCALE DIARIZATION SYSTEM

2.1. Overview of the proposed system

Fig. 2 shows a block diagram of our proposed method as opposed to the conventional speaker diarization pipeline. For the embedding extractor, we employ an x-vector in [6, 17]. We replace the segmentation process with a multi-scale segmentation process followed by a neural affinity score fusion (NASF) system, which will be described in the following sections. The NASF module outputs an affinity matrix similar to that in a conventional speaker diarization framework. In our proposed diarization, we employ the clustering method presented in [18].

2.2. Multi-scale segmentation

Our proposed segmentation scheme for each scale is based on the segmentation scheme that appeared in a previous study [7, 17]. Fig. 3 shows how our proposed multi-scale segmentation scheme works. Although many different scale lengths and numbers of scales can be adopted, we employ three different segment lengths: 1.5, 1.0, and 0.5 s. The hop-length is half the segment length, which is 0.75, 0.5, and 0.25 s, respectively. In addition, the minimum segment length of the each scale is set to 0.5, 0.25, and 0.17 s, respectively.

We refer to the finest scale, 0.5 s, as the *base scale* because the unit of clustering and labeling is determined by base scale. For each base scale segment, we select and group the segments from the lower temporal resolution scales (1.0 s and 1.5 s) whose centers are the closest to the center of the base scale segment. This mapping is shown by the red bounding boxes in Fig. 3. By selecting the segments as in Fig. 3, the clustering results are generated based on the base scale segments, whereas measuring the affinity for the clustering process is achieved using the distance obtained from all three scales.

3. NEURAL AFFINITY SCORE FUSION MODEL

For the speaker diarization task, learning an affinity fusion model is not a straightforward downstream task unlike training speaker embedding from speaker labels because the diarization output is obtained through a clustering (unsupervised learning) approach. Thus, we derived an indirect method that can learn a model for estimating the desirable weights for the affinity scores from multiple scales.

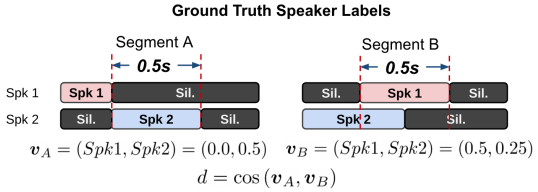

Fig. 4. Example of training data label generation.

3.1. Creating Data Labels

To represent the ground-truth composition of the speakers in the given segments, we employ a concept of a *speaker label vector* based on the duration of each speaker. The dimensions of the speaker label vector are determined based on the total number of speakers in a session. Fig. 4 shows an example of how we create labels of training data. Let segments A and B be a pair of segments for which we want to obtain an affinity score label. In Fig. 4, the speaker label vector v_A obtains values of $(0, 0.5)$ and $(0.5, 0.25)$ from the duration of the speaker labels from segments A and B, respectively.

Because the speaker label vectors are always positive, the ground truth cosine distance value ranges from zero to one. To match the range, the cosine similarity value from the speaker embeddings are min-max normalized to the (0, 1) scale. In total, for L segments in the given session, we obtain ${}_LC_2$ ground truth affinity score labels, which were created for the base scale, which has a segment length of 0.5 s.

3.2. Affinity Score Fusion Networks

To tackle the affinity weighting task, we employ a neural network model optimized using the Mean Square Error (MSE) between the ground truth cosine similarity d and weighted cosine similarity value y. We expect the estimated weight to minimize the gap between the ideal cosine similarity and the weighted sum of the given cosine similarity values (c_1, c_2, c_3). To achieve this, we employ an architecture similar to that of a Siamese network [19], which shares the weights of the networks to process the two different streams of information. Thus, we build a neural network model that can capture the non-linear relationship between a set of affinity weights and a pair of speaker representations by setting up a pair of cloned neural networks.

Fig. 5 shows the architecture of the affinity score fusion network. After the multi-scale segmentation process, embeddings for each scale are extracted with three different embeddings for the three scales. The set of embeddings (segment set A) are then processed using three parallel multi-layer perceptrons (MLPs) and the output of the MLPs is merged to form an embedding from all three scales. The forward propagation of the input layer to the merging layer is also applied to another set of segments (segment set B) to obtain a merged embedding for this set. After forward propagation of two streams of input, the difference between two merged embeddings are passed to the shared linear layer, which outputs the softmax values. We then take the mean of the softmax values from N input pairs.

$$\mathbf{w} = \left(\frac{1}{N}\sum_{n=1}^{N} w_{1,n}, \frac{1}{N}\sum_{n=1}^{N} w_{2,n}, \frac{1}{N}\sum_{n=1}^{N} w_{3,n} \right) \quad (1)$$

The set of averaged softmax values $\mathbf{w} = (\bar{w}_1, \bar{w}_2, \bar{w}_3)$ weights the

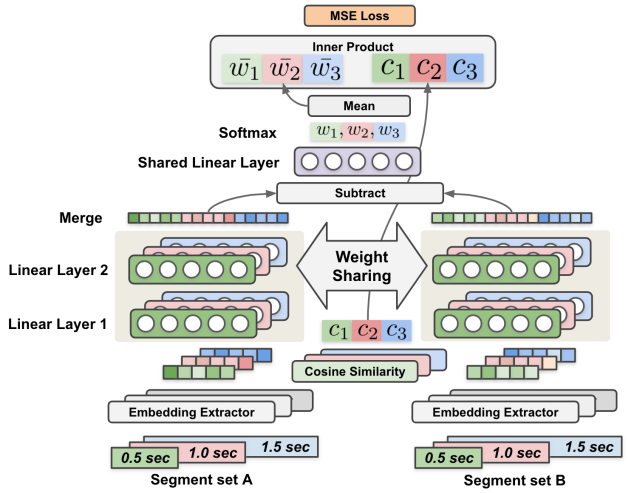

Fig. 5. Neural multi-scale score fusion model

cosine similarity values, $\mathbf{c} = (c_1, c_2, c_3)$, which are calculated using the speaker representations to obtain the weighted cosine similarity value as follows:

$$y_n = \sum_{i=1}^{3} \bar{\omega}_i c_{i,n} = \mathbf{w}^T \mathbf{c}_n, \tag{2}$$

where y_n is the output of the affinity weight network for the n-th pair out of N pairs. Finally, the MSE loss is calculated using the ground truth cosine similarity value d as follows:

$$\mathcal{L}(\mathbf{y}, \mathbf{d}) = \frac{1}{N} \sum_{n=1}^{N} (y_n - d_n)^2, \tag{3}$$

where d_n is the n-th ground-truth cosine score for the n-th pair of segments. In inference mode, we also take the mean of N sets of softmax values to obtain a weight vector $\mathbf{w}$.

3.3. Weighted Sum of Affinity Scores

Our proposed system estimates a weight vector $\mathbf{w}$ for each input session (an independent audio clip under a real-world scenario). For inference of the affinity weight, we randomly select N=$5 \cdot 10^5$ samples out of ${}_L C_2$ pairs per session, which has L base scale segments, and weigh the given affinity matrices as indicated in Fig. 6. The weighted affinity matrix is then passed to the clustering module.

3.4. Affinity Matrix and Clustering

In our previous study [18], we showed that the cosine similarity when applying the NME-SC method can outperform that of the prevailing clustering approaches, such as a probabilistic linear discriminant analysis (PLDA) coupled with agglomerative hierarchical clustering (AHC). Thus, we employ cosine similarity and NME-SC method in [18] to verify the efficacy of the proposed multi-scale affinity weight model by showing the additional improvement from the results in [18]. In addition, we compare the performance with systems based on single-scale segmentation methods.

4. EXPERIMENTAL RESULTS

4.1. Datasets

4.1.1. CALLHOME (NIST SRE 2000)

NIST SRE 2000 (LDC2001S97) is the most widely used diarization evaluation dataset and is referred to as CALLHOME. To compare

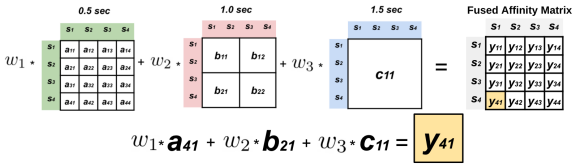

Fig. 6. Example of weighted sum of affinity matrices

its performance with performance of previous studies [6, 17], a 2-fold cross validation is conducted for evaluation on CALLHOME for AHC coupled with the PLDA method.

4.1.2. Call Home American English Speech (CHAES)

CHAES (LDC97S42) is a corpus that contains only English speech data. CHAES is divided into train (80), dev (20), and eval (20) splits.

4.1.3. AMI meeting corpus

The AMI database consists of meeting recordings from multiple sites. We evaluated our proposed systems on the subset of the AMI corpus, which is a commonly used evaluation set that has appeared in numerous previous studies, and we followed the splits (train, dev, and eval) applied in these studies [20–22].

4.2. Training of NASF network

For the training of our proposed neural network model, we use the CHAES- and AMI-train splits. We also apply the CHAES-dev and AMI-dev sets to tune the hyper-parameters of the network. We use MLPs with 2 hidden layers and 128 nodes each and apply the Adam optimizer with a learning rate of 0.001.

4.3. Distance measure and clustering algorithm

In this paper, all systems employ a speaker embedding extractor (x-vector) that appeared in [6, 17]. The following baselines are for the distance measure and clustering method.

4.3.1. PLDA+AHC

This approach is based on the AHC algorithm coupled with the PLDA as it appeared in [6, 17]. The stopping criterion of the AHC was selected based on a grid-search for each development set. We use the PLDA model provided in [17].

4.3.2. COS+NME-SC

As stated in [18], NME-SC does not require a development set to tune the clustering algorithm. We use the same set of segments and speaker embeddings as PLDA+AHC but replace the distance measure with NASF from three different scales and replace the clustering algorithm with NME-SC. In this study, we do not evaluate combinations such as PLDA+NME-SC or COS+AHC because such combinations of algorithms have under-performed PLDA+AHC and COS+NME-SC [23].

4.4. Diarization evaluation

4.4.1. Inference Setup

- Equal Weight: This system is evaluated to show the efficacy of the NASF method over naive cosine similarity averaging. An equal weight system does not use any inference and applies equal affinity weights $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ for all sessions in all datasets.
- NASF-D: This system divides the input session into three equal-length sub-sessions and estimates six different affinity weight vectors ($\mathbf{w}$) for the six different affinity matrices ($3! = 6$), which are intra-sub-session (three sessions) and inter-sub-session (three sessions) affinity matrices. Finally, we calculate the weighted sum

Table 1. Experimental results of baselines and the proposed methods

Dataset	Number of Sessions	PLDA+AHC			Previous Studies	COS+NME-SC			Multi-scale COS+NME-SC		
		$0.5s$	$1.0s$	$1.5s$		$0.5s$	$1.0s$	$1.5s$	Equal Weight	NASF-D	NASF-S
AMI	12	38.42	20.07	10.55	8.92 [20]	26.96	9.82	3.37	6.51	3.89	**3.32**
CHAES-eval	20	4.58	3.15	3.28	2.48 [18]	8.71	3.35	2.48	2.52	2.47	**2.04**
CALLHOME	500	17.89	9.13	8.39	6.63 [24]	20.96	7.81	7.29	6.64	7.02	**6.46**

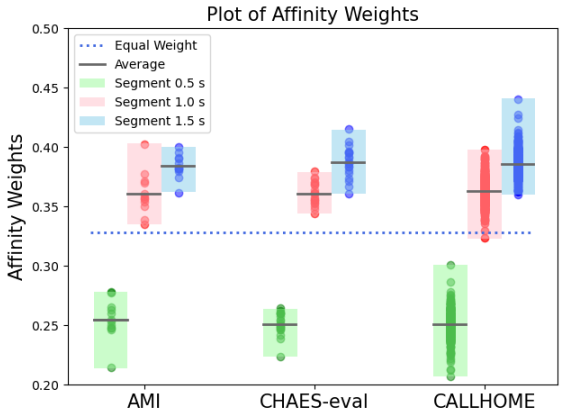

Fig. 7. Plot of affinity weights by datasets.

of these matrices and join the affinity matrices to cluster the integrated matrix as a single affinity matrix.

- NASF-S: This system estimates a set of affinity weights for an entire session. Thus, we have one affinity weight vector **w** for each session, and the entire affinity matrix is weighted using this single weight vector.

4.4.2. DER calculation

To gauge the performance of speaker diarization accuracy, we use oracle SAD output that excludes the effect of SAD module. For all evaluations and datasets, we estimate the number of speakers in the given session without additional information about speaker numbers. We employ an evaluation scheme and software that appeared in [25] to calculate Diarization Error Rate (DER).

4.5. Discussions

We compare the DER obtained from the proposed method with the DER values obtained from each segment scale. Table 1 shows the DER from numerous settings and datasets. We show the DER values of the PLDA+AHC approach for three different segment scales (1.5, 1.0, and 0.5 s) and how the performance of the diarization changes with the distance measure and clustering method. We also list the lowest DER value that we could find that has appeared in a published paper on speaker diarization [20, 24], including the CHAES-eval results of our previous study [18].

Most importantly, we compare the COS+NME-SC methods with segment lengths of 0.5, 1.0, and 1.5 s with the proposed method. The best performing system, NASF-S, obtains relative improvements with error rates of 1.5%, 17.3%, and 11.4% for AMI, CHAES-eval, and CALLHOME, respectively, over the 1.5-s COS+NME-SC baseline. For the AMI corpus, the improvement was minor whereas the CALLHOME and CHAES-eval sets showed a significant improvement given that the DER result from COS-NME-SC with 1.5-s segments is already competitive compared to the results appearing in previous published studies. In Fig. 7, we show the ranges and averages of the estimated weights over the sessions in each dataset. We can see that only the CALLHOME dataset shows a range that includes equal weight within the weight range for 1.0-s segments, whereas the weight ranges from AMI and CHAES-eval show no overlap with an equal weight. We conjecture that this is related to the result in which an equal weight shows an improvement for only CALLHOME.

From the experiment results, we can obtain a few valuable insights. The equal weight experiment gives conflicting results for AMI and CALLHOME. Nevertheless, from the equal weight experiment, we can verify that the desirable affinity weight cannot be simply found by averaging it and that the NASF approach can be a solution for estimating the desirable weights.

The difference in performance gains between AMI and CALLHOME also shows the characteristics of a multi-scale approach. Because the longest segment we employ in our system is 1.5 s, we can argue that the DER reduction comes from the higher resolution of the segments. This becomes clear if we compare the proposed method with the DER we obtain from 0.5-s segments. However, the gain from our proposed method was not that significant with the AMI corpus. We speculate that this is caused by the characteristics of the dataset because the average length of the continuous speaker homogeneous region in the AMI corpus is 2.56 s, whereas the lengths for CALLHOME and CHAES are 2.17 and 2.07 s, respectively. In this sense, we can argue that the CALLHOME and CHAES datasets are more likely to benefit from the proposed multi-scale diarization approach because a higher resolution can capture shorter speaker homogeneous regions.

Another important finding obtained from this study is that varying the affinity weights in a session (i.e., a diarization session that is being clustered) does not lead to a good performance. Having a constant affinity weight in a single affinity matrix leads to a better performance, as we can see from the NASF-S outperforming NASF-D.

5. CONCLUSIONS

We proposed a method that mitigates the limit of the trade-off between the temporal resolution and the fidelity of the speaker representations. The proposed method estimates a set of weights that minimizes the gap between the weighted sum of the cosine affinity and the ground-truth affinity between a pair of segments. The proposed NASF system has a temporal resolution of 0.5 s and improves the diarization performance over conventional single-scale systems, achieving a state-of-the-art performance on the CALLHOME dataset. We believe that the proposed multi-scale score fusion approach on a diarization task can achieve a breakthrough in such research.

Further studies will include an online version of multi-scale speaker diarization where we can find the weights of the affinities in an online fashion. We expect our proposed multi-scale diarization framework to be applicable to many different diarization studies because our method is compatible with other modules in the diarization pipeline.

6. REFERENCES

[1] 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.

[2] Scott Chen, Ponani Gopalakrishnan, et al., "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA broadcast news transcription and understanding workshop*. Virginia, USA, 1998, vol. 8, pp. 127–132.

[3] Matthew A Siegler, Uday Jain, Bhiksha Raj, and Richard M Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA speech recognition workshop*, 1997, vol. 1997.

[4] Stephen H Shum, Najim Dehak, Réda Dehak, and James R Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2015–2028, May 2013.

[5] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 413–417.

[6] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2018, pp. 5329–5333.

[7] Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, et al., "Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge.," in *Proc. INTERSPEECH*, Sep. 2018, pp. 2808–2812.

[8] Mohammed Senoussaoui, Patrick Kenny, Themos Stafylakis, and Pierre Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–227, 2013.

[9] Federico Landini, Shuai Wang, Mireia Diez, Lukáš Burget, Pavel Matějka, Kateřina Žmolíková, Ladislav Mošner, Oldřich Plchot, Ondřej Novotnỳ, Hossein Zeinali, et al., "But system description for dihard speech diarization challenge 2019," *arXiv preprint arXiv:1910.08847*, 2019.

[10] Youngmoon Jung, Seongmin Kye, Yeunju Choi, Myunghun Jung, and Hoirin Kim, "Multi-scale aggregation using feature pyramid module for text-independent speaker verification," *arXiv preprint arXiv:2004.03194*, 2020.

[11] Yun Tang, Guohong Ding, Jing Huang, Xiaodong He, and Bowen Zhou, "Deep speaker embedding learning with multi-level pooling for text-independent speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6116–6120.

[12] MAH Huijbregts, David A van Leeuwen, and FM Jong, "The majority wins: a method for combining speaker diarization systems," *Proc. INTERSPEECH*, 2009.

[13] Simon Bozonnet, Nicholas Evans, Xavier Anguera, Oriol Vinyals, Gerald Friedland, and Corinne Fredouille, "System output combination for improved speaker diarization," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[14] Andreas Stolcke and Takuya Yoshioka, "Dover: A method for combining diarization outputs," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 757–763.

[15] Jose Pardo, Xavier Anguera, and Chuck Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1212–1224, 2007.

[16] Bing Yin, Jun Du, Lei Sun, Xueyang Zhang, Shan He, Zhenhua Ling, Guoping Hu, and Wu Guo, "An analysis of speaker diarization fusion methods for the first dihard challenge," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1473–1477.

[17] David Snyder, "Callhome diarization recipe using x-vectors," Github, May 4, 2018. [Online]. Available: `https://david-ryan-snyder.github.io/2018/05/04/model_callhome_diarization_v2.html`, [Accessed Feb. 11, 2021].

[18] Tae Jin Park, Kyu J Han, Manoj Kumar, and Shrikanth Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2019.

[19] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proceedings of the Workshop on Deep Learning in International Conference on Machine Learning, ICML*, 2015.

[20] Monisankha Pal, Manoj Kumar, Raghuveer Peri, Tae Jin Park, So Hyun Kim, Catherine Lord, Somer Bishop, and Shrikanth Narayanan, "Speaker diarization using latent space clustering in generative adversarial network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6504–6508.

[21] Guangzhi Sun, Chao Zhang, and Philip C Woodland, "Speaker diarisation using 2d self-attentive combination of embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5801–5805.

[22] Sree Harsha Yella and Andreas Stolcke, "A comparison of neural network feature transforms for speaker diarization," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[23] Tae Jin Park, Manoj Kumar, Nikolaos Flemotomos, Monisankha Pal, Raghuveer Peri, Rimita Lahiri, Panayiotis Georgiou, and Shrikanth Narayanan, "The second dihard challenge: System description for usc-sail team," *Proc. INTERSPEECH*, pp. 998–1002, 2019.

[24] Qingjian Lin, Ruiqing Yin, Ming Li, Hervé Bredin, and Claude Barras, "Lstm based similarity measurement with spectral clustering for speaker diarization," 2019, pp. 366–370.

[25] Jonathan G Fiscus, Jerome Ajot, Martial Michel, and John S Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *Proc. Int. Workshop Mach. Learn. Multimodal Interaction*, May 2006, pp. 309–322.