SELF-SUPERVISED LEARNING FOR AUDIO-VISUAL SPEAKER DIARIZATION

[1]*Yifan Ding,* [2]*Yong Xu,* [2]*Shi-Xiong Zhang,* [3]*Yahuan Cong and* [1]*Liqiang Wang*

[1]University of Central Florida, USA, [2]Tencent AI Lab, USA, [3]BUPT, China

ABSTRACT

Speaker diarization, which is to find the speech segments of specific speakers, has been widely used in human-centered applications such as video conferences or human-computer interaction systems. In this paper, we propose a self-supervised audio-video synchronization learning method to address the problem of speaker diarization without massive labeling effort. We improve the previous approaches by introducing two new loss functions: the dynamic triplet loss and the multinomial loss. We test them on a real-world human-computer interaction system and the results show our best model yields a remarkable gain of +8% F_1*-scores* as well as diarization error rate reduction. Finally, we introduce a new large scale audio-video corpus designed to fill the vacancy of audio-video dataset in Chinese.

Index Terms— Speaker diarization, multi-modal learning, self-supervised learning, audio-video synchronization

1. INTRODUCTION

Speaker diarization is the process of partitioning an input audio or video stream into individual segments to match specific speakers. It is one of the core components for many human-centered applications such as video conference systems, human-robot or human-computer interactions, and video re-targeting problems [1, 2, 3]. For example, in a human-computer interaction system, multiple people may talk to the system simultaneously, and we need to identify individual active speakers and separate their faces/bodies and audios before analyzing their activities.

The diarization can be performed on video, audio, or both. Many studies focus on either video-only or audio-only approaches. Everingham et al. use the movement of lips (*i.e.*, video only) to define active speakers [4]. While in most cases, only audio is used [2, 3, 5, 6]. Recently, multi-modal (*e.g.*, audio-video) approaches are attracting more attention [7, 8, 9]. Supervised approaches have been proposed to identify speakers based on the correlation between the audio and video features [10, 11, 12], which requires per frame labeling. To relieve people from massive labeling work, unsupervised or self-supervised methods are proposed [13, 14]. Chung et al. suggest an end-to-end audio-video synchronisation system to use synchronisation between video and audio as the supervision signal. In the approach, a 3D convolutional network is designed to extract video features and contrastive loss is applied [15]. Recently, Nagrani presents a method for speaker identification/verification [16]. In 2018, Korbar et al. presents a self-supervised method is proposed to use curriculum learning and delicate negative example selection strategy, where contrastive loss is employed and claimed as a critical component [17].

However, a disadvantage of contrastive loss is that all unsynchronized pairs are treated equally as negative pairs. Specifically, an audio-video pair with only one frame shift from each other is treated the same as a heterologous pair where the audio and video are from different sources, causing a serious imbalance between positive and negative samples. Besides, human can barely detect lip-sync errors below 200 ms and the training videos downloaded online are sometimes unsynchronized for a few frames due to recording or uploading errors. It harms the model performance to treat these slightly unsynchronized audio-video pairs equally with largely shifted pairs or heterologous audio/video pairs. To relief this problem, Chung et al. tried the classification (*i.e.*, cross entropy) loss but was unable to achieve convergence [15].

In this paper, we first propose to sample three kinds of audio-video pairs for training: synchronized pairs, shifted pairs in which audio and video are shifted by j video frames, and heterologous pairs where audio and video belong to different sources. Then it comes to our major contributions - to propose two new loss functions: dynamic triplet loss and multinomial loss. Like standard triplet loss, the distance between negative pairs should exceeds the positive pairs by a margin. The difference is that in dynamic triplet loss, positive pairs and negatives pairs are dynamically determined in each iteration. Our experiment shows that it achieves better performance and converges faster than contrastive loss. But it is still slow because in each iteration, only one positive pair and one negative pair are sampled. It takes many iterations to sample all shifting combinations of each audio-video segment, which makes it harder to find the global optimum.

To solve the above problems, we further propose the multinomial loss, where all shifting combinations for an audio-video pair and all heterologous combinations for audio-

This works was done while Yifan Ding and Yahuan Cong were interns in Tencent AI Lab, Bellevue, USA.

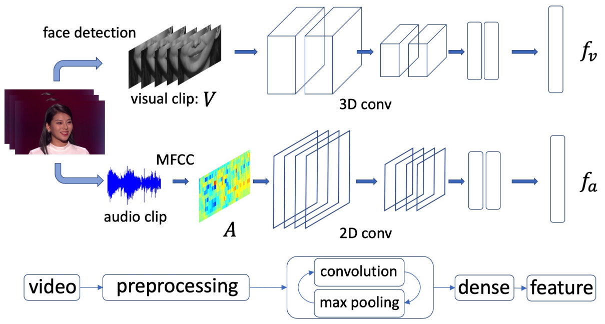


Fig. 1. Two stream network architecture.

video pairs within a mini-batch are considered simultaneously. Specifically, we cluster the negative pairs into groups, where different margins with LogSumExp (LSE) are employed to achieve a smooth maximum [18] in each group. The experiment results show that multinomial loss achieves even faster convergence and better performance compared with dynamic triplet loss and contrastive loss.

Finally, we propose a new large scale audio-video corpus in Chinese [19] to fill the vacancy of such kind of training data. All our experiments are tested on a real-world human-computer interaction system and the results show the effectiveness of our proposed method.

2. OUR METHOD

To process multi-modal signals, we use a two-stream network to extract audio and video features separately. The proposed dynamic triplet loss and multinomial loss are employed to optimize the network.

2.1. Two-stream network structure

To achieve speech diarizaiton, we process audio and video separately [15], which is a common approach for multi-modal tasks. For audio stream, the input is first transformed to MFCC (Mel Frequency Cepstral Coefficient) features, *i.e.*, a power spectrum of a sound on a non-linear mel scale of frequency. Then the MFCC is sent to a 2D convolutional network to produce speech feature. For video stream, a 3D convolution module is employed to extract both temporal information between consecutive video frames and spatial information in each video frame. We use f_a and f_v to denote the audio and video streams, respectively. Figure 1 shows the structure.

2.2. Sampling strategy

Suppose we have a visual segment V_n, where $n \in \{1, 2...N\}$, N is the total number of visual segments. Then we define its correspondent synchronized audio segment as A_n^S. A shifted audio segment from the same video but with j shifted frame is denoted by A_n^j, where $j \in \{-T, ... -1, 1...T-1, T\}$, T is the pre-defined shifting range, which is 10 in our experiment. Let A_n^H denote a heterologous audio segment from another video. Specifically, we consider three types of audio-video pairs: synchronized pairs (V, A^S), shifted pairs (V, A^j), and heterologous pairs (V, A^H). All visual and audio segments have consistent length (5 video frames in our experiment). The sampled pairs are demonstrated in Figure 2.

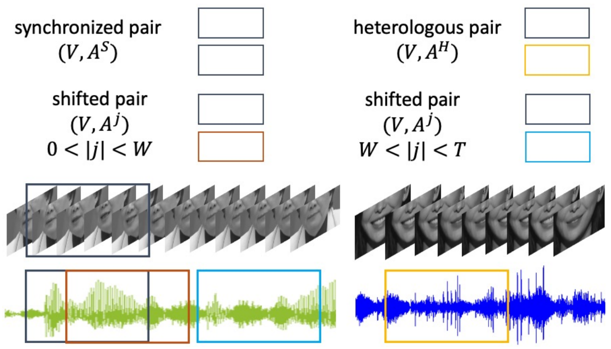


Fig. 2. Examples of audio-video synchronized, shifted, and heterologous pairs. W denotes the lenght of visual clip. T is the shifting range.

2.3. Dynamic triplet loss

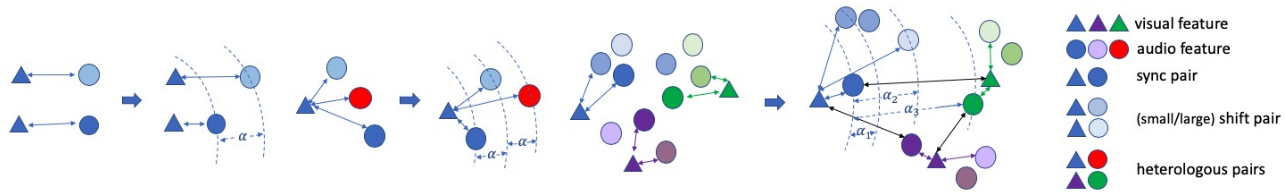
In contrastive loss, the distance between unsynchronized audio-video pairs are pushed larger than synchronized audio-video pairs. The loss is as follows:

$$L_{con} = \frac{1}{2N} \sum_{n=1}^{N} (y_n) d_n^2 + (1 - y_n) \max(\alpha - d_n, 0)^2 \quad (1)$$

$$d_n = ||f_v(V_n) - f_a(A_n)||_2 \quad (2)$$

where $y \in [0, 1]$ is the binary similarity metric denoting whether the visual and audio segments are synchronized, and $y = 1$ means synchronized.

A problem in the contrastive loss is that it equally treats all unsychronized pairs. Since negative pairs are sampled by shifting the audio/visual segment or replacing the audio/visual clips from another video. Serious imbalance lies between the number of positive and negative pairs. In other words, there are much more shifted and heterologous pairs than positive synchronized pairs. In addition, the model would more likely be dominated by easy negatives (*e.g.*(V_n, A_n^H) from different videos) rather than hard negatives (*e.g.*(V_n, A_n^j) with small shifts). Finally, the training samples downloaded online could be slightly unsynchronized due to possible recording/uploading error, which may also cause the algorithm to diverge.



(a) Contrastive loss (b) Dynamic triplet loss (c) Multinomial loss

Fig. 3. Comparison of loss functions. **(a):** The negative pair is separated by a margin. **(b):** Negative pairs are separated with each other by a margin. **(c):** Negative pairs with different shifts or heterologous audio are separated by separate margins.

In this paper, we propose dynamic triplet loss together with ,.the aforementioned sampling strategy to solve these problems. Specifically, we first sample audio-video training data using the three sampling methods introduced in Section 2.2 to obtain synchronized pairs (V_n, A_n^S), shifted pairs (V_n, A_n^j), and heterologous pairs (V_n, A_n^H). During the training, the positive and negative pairs are dynamically defined according to their relative distance, as shown in Table 1, *i.e.*, a positive pair in one iteration could be a negative pair in another iteration. Finally, the model is optimized with Equation 3.

$$L_{D\_tri} = \sum_{n=1}^{N}[||f_v(V_n) - f_a(A'_n)||_2^2 - ||f_v(V_n) - f_a(A''_n)||_2^2 + \alpha]_+ \tag{3}$$

where (V, A') refers to a positive pair, (V, A'') refers to a negative pair, and α is the pre-defined margin. By using the dynamic triplet loss, the algorithm progressively learns the following distance rule: $D(f_v(V_n), f_a(A_n^S)) < D(f_v(V_n), f_a(A_n^{j}{}')) < D(f_v(V_n), f_a(A_n^{j}{}'')) < D(f_v(V_n), f_a(A_n^H))$, where $j', j'' \in \{-T, ... - 1, 1...T - 1, T\}$ and $|j'| < |j''|$. D denotes the distance function measured by l_2. f_a and f_v denote the audio and visual feature extraction functions, respectively.

Table 1. Definition of positive and negative pairs according to their relative distance.

Index	Positives (V_n, A'_n)	Negatives (V_n, A''_n)
Case-1	(V, A^S)	(V, A^j)
Case-2	$(V, A^{j'})$	$(V, A^{j''})(\lvert j'\rvert < \lvert j''\rvert)$
Case-3	(V, A^j)	(V, A^H)

2.4. Multinomial loss

In dynamic triplet loss, only two pairs are sampled in each iteration. However, for each audio-video pair, there are C_{2T}^2 shifting options and more heterologous options. To best take advantage of these combinations during each training iteration, we further propose an improved loss called *multinomial loss* by optimizing the positive pairs and negative pairs by clusters, while each cluster has its own optimization margin. In this way, the negative pairs with different specialities (*e.g.*shifting range) would be treated differently. We propose to cluster the negative pairs into groups and apply separate margins for each group. Equation 4 defines the loss.

$$L_{mul} = \sum_{n=1}^{N}(D(f_v(V_n), f_a(A_n^S)) + \sum_{k}^{K}\log(\sum_{u}^{u\in cluster\{k\}} \exp(\alpha_k - D(f_v(V_n), f_a(A_n^u)))) \tag{4}$$

$$L_{mul} = \sum_{n=1}^{N}(D(f_v(V_n), f_a(A_n^S)) + \log(\sum_{j'}^{0<|j'|\leq m_1} \exp(\alpha_1 - D(f_v(V_n), f_a(A_n^{j'})))) + \log(\sum_{j''}^{m_1<|j''|\leq m_2} \exp(\alpha_2 - D(f_v(V_n), f_a(A_n^{j''})))) + \log(\sum_{H}^{H\neq n} \exp(\alpha_3 - D(f_v(V_n), f_a(A_n^H))))) \tag{5}$$

Specifically, in our audio-video synchronization learning method, we separate our loss into four parts, which is shown as Equation 5. The first part is to minimize the distance of the synchronized pairs $D(f_v(V_n), f_a(A_n^S))$. The second part denotes the loss for shifted pairs when the shifting distance is within m_1. We set m_1 to be equal to the size of video segment, which is 5 frames in our experiment. *LogSumExp* [18] is applied to achieve a smooth maximum and α_1 is a margin to this loss. The third and fourth loss functions are similar but for different audio-video pairs. The third loss is for shifted audio-video pairs where the shifting range from m_1 to m_2, when the shifted audio and visual segments are from the same video but not temporally overlapped. We set m_2 to be 10 frames and α_2 is the corresponding margin. The forth loss and margin α_3 is for heterologous pairs, *i.e.*, visual and audio segments from different videos in the mini-batch.

3. EXPERIMENTS

In this section, we describe the dataset we uses, the test metrics and the experimental results.

3.1. Dataset and baselines

Our training dataset contains over 140 thousand video clips, about 100 hours long in total. All video clips are from YouTube with the front face facing the camera. The test-set includes 12 videos. Each is 40-60s long and contains 3 to 4 people talking in turn. Our data would be described in details and released soon in another paper [19]. We compare our results with SyncNet [15] and UIS-RNN (unbounded interleaved-state recurrent neural networks) [5]. UIS-RNN is a fully-supervised audio-only speaker diarization system which takes d-vector embedding as input and each individual speaker is modeled by a parameter-sharing RNN, while the RNN states for different speakers interleave in the time domain. We also extended the UIS-RNN [5] with the number of detected faces as the interleaved-state upbound in RNN. This means the number of faces in the video are used to indicate the number of potential speakers.

3.2. Testing results

Table 2. Model Comparison (%))

#	Method	F_1-scores	*DER*
1	SyncNet [15]	76.4	23.1
2	UIS-RNN [5]	70.0	25.6
3	face-bounded UIS-RNN	72.6	22.2
4	ours-triplet	75.3	22.9
5	ours-triplet-LipNet[20]	76.8	21.7
6	ours-cluster	**84.9**	**17.0**

In this section, we compare our results with the baselines. For all the models, the faces are detected with Dlib [21]. Gray-scale images of the lower half face is resized to be 112 × 112 for training. For the audio part, a 13 mel MFCC feature are used in our methods. We use 25FPS for the visual clips and 100Hz for the audio segments. Therefore the length of visual and audio input is 5-frame and 20-frame respectively. The value of $\alpha_1, \alpha_2 and \alpha_3$ are set to be 1,2 and 10 respectively. A batch-size of 16 is used for all experiments and no data augmentation is implemented. Besides, the distances are generated through a per-frame evaluation of the l_2 distance between the audio and visual feature for each speaker in the video. The speaker (*i.e.*face) feature which has the lowest distance with the audio feature would be identified as the active speaker. Table 2 shows the results.

We use *DER* (Diarization Error Rate) and F_1-*scores* to evaluate the performance of each model. Compared with SyncNet [15], our best model improved 8% and 6% on F_1-*scores* and *DER*, respectively. To make it clear, the same MLP network is employed in model #1 (“SyncNet”) and model #6 (“ours-cluster”). However we find incorporating LipNet [20] is also beneficial to the performance since LipNet has a more complex architecture (*i.e.*ResNet [22]) compared with 6-layer MLP (multi layer perceptron) we used in model #4 (“ours-triplet” in Table 2). Figure 4 shows the visualization of per-frame audio-video distances generated by our proposed multinomial model (“(a)”) and SyncNet (“(b)”). From which we can find: for our model, the distance of the active speaker is significantly below the distance of non-active speakers. While for SyncNet, the distances of different speaker could hardly be distinguished. While in More video demos can be found from the project homepage.[1]

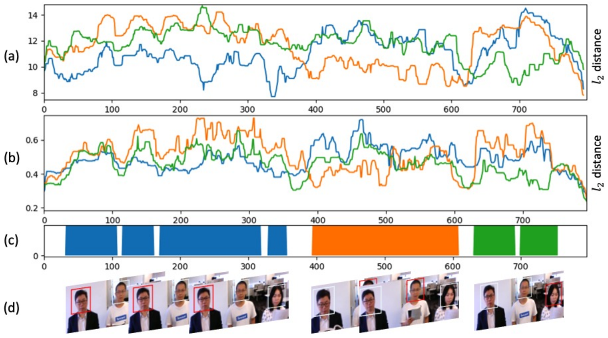


Fig. 4. Per-frame test distances. **(a):** Ours. Different colors denote the a-v distance of different speakers. The curve with lowest distance is the predicted active speaker. **(b):** SyncNet. **(c):** GT. **(d):** Visualization of GT. The red box frames the face of active speakers.

4. CONCLUSION

In this paper, we propose two new losses: the dynamic triplet loss and multinomial loss, and a large scale dataset in Chinese for self-supervised audio-video synchronization learning. The work can benefit the task of speaker diarization, which is an important basic task for many human-centered applications. We demonstrate experiments on a real-world human-computer interaction system and compare our results with several baselines. The results show the proposed methods outperforms both the audio-only and multi-modal previous approaches.

[1] https://yifan16.github.io/av-spk-diarization

5. REFERENCES

[1] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[2] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. 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.

[3] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5239–5243.

[4] M. Everingham, J. Sivic, and A. Zisserman, ""Hello! My name is... Buffy" – automatic naming of characters in TV video," in *British Machine Vision Conference*, 2006.

[5] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6301–6305.

[6] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

[7] K. Hoover, S. Chaudhuri, C. Pantofaru, M. Slaney, and I. Sturdy, "Putting a face to the voice: Fusing audio and visual signals across a video to determine speakers," *arXiv preprint arXiv:1706.00079*, 2017.

[8] E. Marcheret, G. Potamianos, J. Vopicka, and V. Goel, "Detecting audio-visual synchrony using deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] I. D. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1086–1099, 2017.

[10] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep multimodal speaker naming," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1107–1110.

[11] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, listen and learna multimodal lstm for speaker identification," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[12] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi *et al.*, "Ava-activespeaker: An audio-visual dataset for active speaker detection," *arXiv preprint arXiv:1901.01342*, 2019.

[13] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[14] G. Sell and D. 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.

[15] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lipreading, ACCV*, 2016.

[16] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[17] B. Korbar, D. Tran, and L. Torresani, "Co-training of audio and video representations from self-supervised temporal synchronization," *arXiv preprint arXiv:1807.00230*, vol. 3, 2018.

[18] F. Nielsen and K. Sun, "Guaranteed bounds on the kullback–leibler divergence of univariate mixtures," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1543–1546, 2016.

[19] "A large-scale audio-visual corpus for multimodal speaker diarization, speech separation and recognition," in *Preparation*.

[20] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.

[21] D. E. King, "A toolkit for making real world machine learning and data analysis applications in C++." [Online]. Available: https://github.com/davisking/dlib

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.