IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

ViVoLAB Multimodal Diarization System for RTVE 2020 Challenge

Victoria Mingote, Ignacio Viñals, Pablo Gimeno, Antonio Miguel, Alfonso Ortega, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{vmingote, ivinalsb, pablogj, amiguel, ortega, lleida}@unizar.es

Abstract

This paper describes a post-evaluation analysis of the system developed by ViVoLAB research group for the IberSPEECH-RTVE 2020 Multimodal Diarization (MD) Challenge. This challenge is focused on the study of multimodal systems for the diarization of audiovisual files and the assignment of an identity to each segment. In this work, we have implemented two different subsystems to address this task using the images and the audio from files separately. To develop our subsystems, we have employed the state of the art speaker and face verification embeddings extracted from publicly available Deep Neural Networks (DNN). Different clustering approaches are also used in combination with the tracking and identity assignment process. Furthermore, in the face verification system, we have included a novel approach to train an enrollment model for each identity which we have shown previously to improve the results compared to the average of the enrollment data. Using this approach, we train a learnable vector to represent each enrollment character.

Index Terms: face recognition, speaker recognition, deep neural networks, enrollment models, spectral clustering, video processing

1. Introduction

Multimodal biometric verification field consists of identifying people using audiovisual resources. In recent years, this field has been widely investigated due to its great interest which is motivated by the fact that human perception uses not only acoustic information but also visual information to reduce speech uncertainty. Furthermore, this task had been rarely addressed for uncontrolled data due to the lack of this kind of datasets. However, in recent years, several challenges focused on this topic have been developed [1, 2, 3], and also a large amount of multimedia and broadcast data is being produced currently like news, talk shows, debates or series. Therefore, to develop a multimodal biometric system, different tools are required to process these data, detect the presence of people and address the identification of who is appearing and speaking. The approach employed in this kind of systems is known as multimodal diarization.

Many studies focus on the simplest way to perform the multimodal diarization based on having separate systems for speaker and face diarization [3, 4]. Speaker diarization is a widespread task [5, 6] due to its usefulness as pre-processing for other speaker tasks. At the same time, it is still a challenging task because there is no prior information about the number and the identity of speakers in the audio files, and the domain mismatch between different scenarios can produce some difficulties. On the other hand, face diarization has been widely employed as a video indexing tool, and the previous step for face verification [7, 8]. However, in unconstrained videos of real-world scenarios, face images often can appear with large variations, so this kind of system has also found some problems in real-world scenarios. For these reasons, a straightforward score level fusion is usually employed to join the information of both types of systems.

The IberSPEECH-RTVE 2020 Challenges aims to benchmark and further analyze this different kind of diarization systems. With this purpose, two types of diarization evaluations are included in this challenge, Speaker Diarization and Identity Assignment (SDIA) [9], and a Multimodal Diarization (MD) [10]. The former is the most extended kind of diarization combined with the speaker assignment, while the latter combines the previous one with face diarization, which is obtaining more relevance in recent times. Thus, we have focused on this second challenge, and specially we will remark the characteristics of face diarization subsystem.

This paper presents the ViVoLAB system submitted to the IberSPEECH-RTVE 2020 Challenge in MD task. This challenge is focused on segmenting broadcast audiovisual documents and assigning to the segments an identity from a closed set of different faces and speakers. For the challenge, we have processed video and audio tracks independently in order to separately improve their performance. However, the pipeline is very similar in both cases where the differences are the exact approach used in each part of the process. Therefore, initially, the video and audio files are processed. After that, an embedding extractor is used to extract the representations, and finally, clustering and assignment process is applied. To carry out the assignment process in the face subsystem, a new approach based on [11] has been applied to model the enrollment identities. This approach was shown as a promising technique to characterize each enrollment identity with only one learnable vector for the speaker verification task, but this is the first time that this technique has been applied in face verification.

The remainder of this paper is laid out as follows. Section 2 provides a description of the challenge and the dataset employed. In Section 3, we describe the face diarization subsystem. The speaker diarization employed is explained in Section 4. Finally, Section 5 presents and discusses results, and Section 6 concludes the paper.

2. RTVE 2020 Challenge

The RTVE 2020 Challenge is part of the 2020 edition of the Albayzin evaluations [12, 10]. This dataset is a collection of several broadcast TV shows in Spanish language and covering different scenarios. To carry out this challenge, the database provides around 40 hours of shows from the public Spanish Television (RTVE). The development subset of the RTVE2020 database contains two of the parts of the RTVE 2018 database ($dev2$ and $test$ partitions) which are formed by four shows of around 6 hours. Furthermore, this subset also contains a new development partition with nine shows of around 4 hours. The evaluation set consists of fifty-four video files of around 29 hours in total with speaker and face timestamps. Enrollment

10.21437/IberSPEECH.2021-16

data is also provided for 161 characters with 10 pictures and a 20-second video of each character.

3. Face Subsystem

This section describes the different blocks of the face system, including video processing, embedding extraction, training face enrollment models, clustering, tracking, and identity assignment scoring. The block diagram of the face system is depicted in Fig.1.

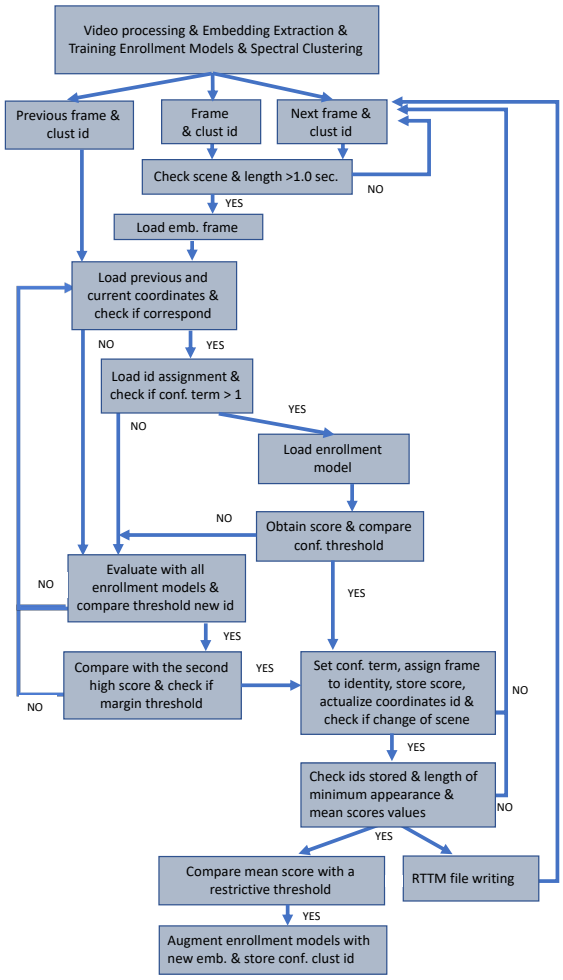

Figure 1: *Block diagram of face system.*

3.1. Video Processing

3.1.1. Frame Extraction

As the first step, we process the video to extract five frames per second using $ffmpeg$ tool [1]. We decided to use five frames per second since this number of frames allows us to have a high precision to determine the limits of the characters appearance.

3.1.2. Face Detector

Face detection is a fundamental step because failures in this process could be crucial for correct development in other parts of the face diarization system. In our system, the face detector employed is a system of alignment and detection based on a deep neural network (DNN) which is called Multi-task Cascaded Convolutional Networks (MTCCN) [13]. In this part, we used this implemented system since it is an effective and contrasted method for face detection, which is necessary to do before continuing with the rest of the face verification pipeline. Furthermore, using this detector, we can store the bounding boxes created by the algorithm which correspond to the coordinates where a face is detected, and we use this information in the tracking and identity assignment processes.

3.1.3. Change Shot Detection

The type of videos employed in this challenge are obtained from television programs, so these programs are usually composed of a huge variability in the content characteristics and constant changes of shot and scenes. Thus, to help the tracking and clustering step, we use a scene detection tool [2] which detects effectively these changes using the threshold-based detection mode. This detector finds areas where the difference between two subsequent frames exceeds a threshold value.

3.2. Embedding Extraction

Once the video processing step is done, we process the face images using the bounding boxes, apply mean and variance normalization, and resize to 160×160 pixels applying a central cropping. After that, the processed images are passed through a trained model to obtain embedding representations. In this system, as a face extractor, we have employed a pretrained convolutional neural network (CNN) with more than one hundred layers [14, 15]. This network was trained for a classification task on the CASIA-WebFace dataset [16], but the embeddings extracted from it have been proved previously in a verification task to check their discriminative ability with impressive results. For this reason, we decide to use these embeddings of 128 dimensions to extract the representations for enrollment and test files of this challenge.

3.3. Training Face Enrollment Models

Traditionally, in recognition tasks, a back-end is applied to compare enrollment and test embeddings and obtain the final verification scores. A widely used approach is cosine similarity where if an enroll identity has more than one enrollment embedding, these embeddings are averaged to compare with the test embedding. However, we demonstrated in [11] for the speaker verification task that a better solution to make this process consists of training an enrollment model for each enroll identity. Thus, in this work, we have applied this approach for face verification task where we have trained one model for each of the 161 enrollment identities. To train these models, we have used the embeddings of enrollment images, and video files from the development and test sets of the IberSPEECH-RTVE 2020 Challenge [10] as positive examples. While the enrollment files from the development and test sets of the IberSPEECH-RTVE 2018 Challenge [12] are used as negative examples.

Fig.2 shows the process to make this training where a learnable vector is obtained to represent each identity. This process consists of comparing positive or target examples with themselves (s_{tar}), and also with negative or non-target examples (s_{nontar}) using as training objective aDCF loss function [17]

[1] https://www.ffmpeg.org/

[2] https://www.pyscenedetect.readthedocs.io/en/latest/

which is an approximated function of a verification metric. To optimize aDCF loss, the scores used are obtained with cosine similarity.

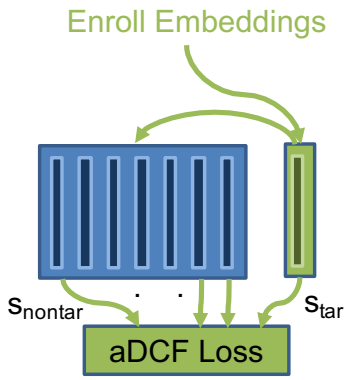

Figure 2: *Training face enrollment models using target and non-target embeddings.*

3.4. Clustering

As a source of complementary information, face embeddings from test videos are used to perform a spectral clustering technique [18] which tries to find strongly connected segments. This technique provides an initial cluster assignation to group the frames in the video sequence. In this work, we have employed this clustering combined with the use of the coordinates to improve the whole tracking process.

3.5. Tracking and Identity Assignment Scoring

Once all the above information is obtained, we have developed an algorithm to carry out the tracking and identity assignment process, which is depicted in Fig.1 and follows a similar philosophy to the one developed in [19]. In this algorithm, the tracking process has been developed by scene, so a shot change restarts the tracking. Therefore, while the scene is the same, the algorithm checks frame by frame the clustering information and the correspondence between the coordinates of the current frame and the previous frame to establish links which allow to make the tracking process. When a relation between both frames exists and has a high confidence term, the identity assignment of the previous frame is used to select the enrollment model and obtain the score. This score is compared with a confidence threshold to determine whether the identity assigned is correct or not. However, when there is no relation between the coordinates of the current frame and the previous frame or the confidence term is low, the frame embedding is compared with all the enrollment models to obtain a score and determine whether is a new identity to assign. Once the identity assignment is made over the current frame, the score is stored, the coordinates are updated, and the algorithm checks whether the scene changes.

Tracking is carried out with the previous steps, but the identity assignment process made is only an initial assignation. When a change of scene is detected, the system checks the identities and scores stored in the scene to remove inconsistent segment assignations. After that, the final segments with their identity assignations are written into the Rich Transcription Time Marked (RTTM) file. In addition, score confidence values are stored when a final identity assignment is made. If these values are greater than a more restrictive threshold which is set with the development set, we augment the enrollment models with the current face embedding. The whole process is repeated with all the detected scenes.

4. Speaker Subsystem

In this section, we present the speaker system which is based on similar blocks to the face system, such as audio processing, embedding extraction, clustering, and identity assignment scoring, but using different approaches in each one.

4.1. Audio Processing

4.1.1. Speech Activity Detection

Our approach for speech activity detection (SAD) is based on a deep learning solution which is an evolution derived from our previous experience with SAD systems in different domains [20]. We use a convolutional recurrent neural network (CRNN) consisting of 3 blocks of 2D convolutional (2 conv layer with 64 filters of size 3x3, batch normalisation and ReLU activation) followed by 3 BiLSTM layers. Then, the final speech score is obtained through a linear layer. As input features, 64 Mel filter banks and the frame energy are extracted from the raw audio and feed to the neural network. Cepstral Mean and Variance Normalization (CMVN) [21] normalization is applied.

4.1.2. Speaker Change Point Detection

The Speaker Change Point Detection block works in terms of Bayesian Information Criterion (BIC), according to its differential form (ΔBIC)[22]. We consider analysis windows of 6 seconds, modelling speakers with full-covariance Gaussian distributions. This block prioritizes those speech/non-speech boundaries given by SAD. As input features, the system considers 20 MFCC [23] features vectors, over a 25 ms hamming window every 10 ms. Features are then normalized according to CMVN to mitigate channel effects.

4.2. Embedding Extraction

Once the audio processing is done, each one of the obtained segments will be transformed into a compact representation also known as embedding. For this purpose, we have opted for an evolution of x-vectors [24] considering an extended version [25] of the TDNN architecture. Compared to the original, we have substituted the original mean and standard deviation pooling block by a multi-head self-attention block [26]. This self-attention block considers H different patterns, also known as heads, learnable from the own data. The output of the block consists of the concatenation of the estimated means and standard deviations. The neural network has been trained with data from VoxCeleb 1 [27] and 2 [28]. The resulting neural network provides embeddings of dimension 512. These embeddings will be later centered, dimensionsality reduced by means of LDA up to 200 and whitening and length-normalized [29].

4.3. Clustering

The obtained embeddings are modeled in a generative manner according to [30], where a tree-based PLDA clustering is proposed. This solution proposes a Maximum A Posteriori (MAP) estimation of the speaker labels Θ given the set of embeddings Φ. The model considers a Fully Bayesian PLDA [31] of dimension 100 to model $P(\Phi|\Theta)$, while the priors [32]. As we

did in [30], we interpret $P(\Phi|\Theta)$ as a tree structure by means of the product rule of probability. Hence, we opt for an optimization of the model according to a sequential manner making use of the M-algorithm [33] to find the best possible path along the tree. Moreover, prior to any clustering evaluation the PLDA model is adapted thanks to unsupervised adaptation approaches as described in [34].

4.4. Identity Assignment Scoring

The Identity Assignment (IA) block follows the schematic of a speaker verification task based on the standard embedding-PLDA paradigm. Hence, as preparation, each one of the enrollment recordings is converted into its corresponding embedding as well as the obtained segments from diarization. For the speaker verification task itself, enrollment models are build according to the correspondings audios while test models represent the clusters obtained during diarization. Each test model is made in terms of all segments assigned to the cluster. For simplicity reasons, we make use of the same embedding extractor and PLDA trained for diarization purposes.

After the scores are obtained, we normalized them using an adaptative s-norm. For each segment, we select cohorts similar to the test segment to compute the normalization values. For each trial, we selected 25% of the total segments in the cohort. The selection is based on the own PLDA scores. The final labels are built according to a threshold adjusted during calibration. This adjustment was obtained experimentally with the development set. Furthermore, as a design choice, we do not exclude the possibility of multiple clusters assigned to the same enrollment. This decision was made as to allow the correction of errors during diarization.

5. Results

In this section, the results for each subsystem are obtained using Diarization Error Rate (DER) as metric to evaluate. DER is usually the reference metric employed in diarization task, but in this case, DER is obtained slightly different than the original metric since it also takes into account the measurement of the identity assignment errors. Table 1 presents DER results obtained in the development and test set for face and speaker modalities. In addition to separate results, we show the mean result achieved with both systems. These results indicate a great mismatch between development and test results. We have analyzed which kind of video files composed both subsets and the length of these files, and we have found that development files are shorter than test files. Thus, we can see that the face and speaker subsystems obtain better performance in development files which are shorter videos, so the tracking process is easier to follow.

Table 1: *Experimental results on RTVE 2020 Multimodal Diariazation set, showing DER%. These results were obtained for the development and test sets in both modalities.*

Subset	Modality	DER%
DEV	FACE	51.66
	SPEAKER	47.90
	FACE+SPEAKER	49.78
TEST	FACE	61.79
	SPEAKER	72.63
	FACE+SPEAKER	67.21

To analyze better these results, Table 2 shows a decomposition of DER metric in the three terms of error:

- *Probability of misses (MISS)*: which indicates the segments where the target identity is presented but the system does not detect it.
- *Probability of false alarm (FA)*: which illustrates the number of errors due to the assignment of one enrollment identity to a segment without identity known.
- *Identity error (ID)*: which reflects the segments assigned to enrollment identities different from the target identity.

Focusing on face modality errors, in the case of development subset, we observe that the main cause of error is the probability of misses which indicates that a huge amount of segments from target identities have not been detected. Therefore, this effect can be motivated by the fact of using a threshold value too high. While in the test subset, misses and false alarm terms are similar. Especially relevant is the great increase of the false alarm errors since this fact illustrates the problems to discard segments of non-target faces when the number of enrollment identities is large. On the other hand, the distribution of errors produced in the speaker subsystem is quite different, because false alarms are much bigger than misses in both subsets of data. Note that it is also related to the threshold chosen. However, in this case, the threshold is lower, so the target segments are mostly detected, but as a result, a high number of enroll identities are assigned to segments of unknown identity.

Table 2: *Decomposition of DER% results in Miss (MISS), False Alarm (FA) and Identity (ID) Errors for the development and test sets in both modalities.*

Modality	Subset	MISS	FA	ID
FACE	DEV	37.5	6.5	7.7
	TEST	29.0	19.5	13.3
SPEAKER	DEV	14.0	29.6	4.3
	TEST	5.1	53.3	14.2

6. Conclusions

This paper presents the ViVoLAB submission to the IberSPEECH-RTVE 2020 Multimodal Diarization Challenge. In this work, we have developed two monomodal subsystems to address separately face and speaker diarization. Each system is based on state-of-the-art DNN approaches. We have demonstrated that there is still room for improvement in each of the systems because the results obtained are too high in both subsets and in both systems. Moreover, future work can be done on the fusion of both systems, which could improve the final results. The high DER values for misses and false alarms in the face and speaker subsystem, respectively, should be addressed by that fusion.

7. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R, by the Government of Aragon (Reference Group T36_20R) and co-financed with Feder 2014-2020 “Building Europe from Aragon”, and by Nuance Communications, Inc. The Titan V used for this research was donated by the NVIDIA Corporation.

8. References

[1] J. Poignant, H. Bredin, and C. Barras, "Multimodal person discovery in broadcast tv at mediaeval 2015," in *MediaEval 2015 working notes proceedings*. CEUR-WS. org, 2015.

[2] H. Bredin, C. Barras, and C. Guinaudeau, "Multimodal person discovery in broadcast TV at MediaEval 2016," in *MediaEval 2016 working notes proceedings*. CEUR-WS. org, 2016.

[3] O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2019 NIST Audio-Visual Speaker Recognition Evaluation," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 259–265.

[4] R. K. Das, R. Tao, J. Yang, W. Rao, C. Yu, and H. Li, "HLT-NUS Submission for NIST 2019 Multimedia Speaker Recognition Evaluation," *arXiv preprint arXiv:2010.03905*, 2020.

[5] 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.

[6] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "In-domain Adaptation Solutions for the RTVE 2018 Diarization Challenge," *Proc. IberSPEECH 2018*, pp. 220–223, 2018.

[7] E. Khoury, P. Gay, and J.-M. Odobez, "Fusing matching and biometric similarity measures for face diarization in video," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, 2013, pp. 97–104.

[8] N. Le, A. Heili, D. Wu, and J.-M. Odobez, "Efficient and Accurate Tracking for Face Diarization via Periodical Detection," in *International Conference on Pattern Recognition*, no. CONF. IEEE, 2016.

[9] A. Ortega, A. Miguel, E. Lleida, V. Bazán, C. Pérez, M. Gómez, and A. de Prada, "Albayzin evaluation: IberSPEECH-RTVE 2020 Speaker Diarization and Identity Assignment," 2020.

[10] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, M. Gómez, and A. de Prada, "Albayzin evaluation: IberSPEECH-RTVE 2020 Multimodal Diarization and Scene Description Challenge," 2020.

[11] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Training Speaker Enrollment Models by Network Optimization," *Proc. Interspeech 2020*, pp. 3810–3814, 2020.

[12] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, and A. de Prada, "Albayzin 2018 evaluation: the iberspeech-RTVE challenge on speech technologies for spanish broadcast media," *Applied Sciences*, vol. 9, no. 24, p. 5412, 2019.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[15] D. Sandberg, "Face Recognition using Tensorflow," https://www.github.com/davidsandberg/facenet.

[16] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[17] V. Mingote, A. Miguel, D. Ribas, A. Ortega, and E. Lleida, "Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems," *Proc. Interspeech 2019*, pp. 2903–2907, 2019.

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation," Tech. Rep., 2000.

[19] E. Ramos-Muguerza, L. Docío-Fernández, and J. L. Alba-Castro, "The GTM-UVIGO System for Audiovisual Diarization," *Proc. IberSPEECH 2018*, pp. 204–207, 2018.

[20] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization Challenge," in *Proc. Interspeech*, 2018, pp. 2803–2807.

[21] M. J. Alam, P. Ouellet, P. Kenny, and D. O'Shaughnessy, "Comparative evaluation of feature normalization techniques for speaker verification," in *International Conference on Nonlinear Speech Processing*. Springer, 2011, pp. 246–253.

[22] S. Chen, P. 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*, vol. 8. Virginia, USA, 1998, pp. 127–132.

[23] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[24] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. 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.

[25] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin *et al.*, "State-of-the-Art Speaker Recognition for Telephone and Video Speech: The JHU-MIT Submission for NIST SRE18." in *Interspeech*, 2019, pp. 1488–1492.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*.

[27] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[28] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[29] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth annual conference of the international speech communication association*, 2011.

[30] I. Viñals, P. Gimeno, A. O. Giménez, A. Miguel, and E. Lleida, "ViVoLAB Speaker Diarization System for the DIHARD 2019 Challenge." in *INTERSPEECH*, 2019, pp. 988–992.

[31] J. Villalba and E. Lleida, "Unsupervised adaptation of plda by using variational bayes methods," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 744–748.

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

[33] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[34] I. Viñals, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Unsupervised adaptation of PLDA models for broadcast diarization," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, pp. 1–13, 2019.