applied sciences

MDPI

Article

Multimodal Diarization Systems by Training Enrollment Models as Identity Representations †

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

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain
* Correspondence: vmingote@unizar.es (V.M.); ivinalsb@unizar.es (I.V.); pablogj@unizar.es (P.G.); amiguel@unizar.es (A.M.); ortega@unizar.es (A.O.); lleida@unizar.es (E.L.)

† This paper is an extended version of our paper published in the conference IberSPEECH2020.

check for updates

Citation: Mingote, V.; Viñals, I.; Gimeno, P.; Miguel, A.; Ortega, A.; Lleida, E. Multimodal Diarization Systems by Training Enrollment Models as Identity Representations. *Appl. Sci.* **2022**, *12*, 1141. https://doi.org/10.3390/app12031141

Academic Editors: Francesc Alías, Valentin Cardeñoso-Payo, David Escudero-Mancebo, César González-Ferreras and António Joaquim da Silva Teixeira

Received: 23 December 2021
Accepted: 19 January 2022
Published: 21 January 2022

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 focuses on the study of multimodal systems for the diarization of audiovisual files and the assignment of an identity to each segment where a person is detected. In this work, we implemented two different subsystems to address this task using the audio and the video from audiovisual files separately. To develop our subsystems, we used the state-of-the-art speaker and face verification embeddings extracted from publicly available deep neural networks (DNN). Different clustering techniques were also employed in combination with the tracking and identity assignment process. Furthermore, we included a novel back-end approach in the face verification subsystem to train an enrollment model for each identity, which we have previously shown to improve the results compared to the average of the enrollment data. Using this approach, we trained a learnable vector to represent each enrollment character. The loss function employed to train this vector was an approximated version of the detection cost function (aDCF) which is inspired by the DCF widely used metric to measure performance in verification tasks. In this paper, we also focused on exploring and analyzing the effect of training this vector with several configurations of this objective loss function. This analysis allows us to assess the impact of the configuration parameters of the loss in the amount and type of errors produced by the system.

Keywords: enrollment models; face recognition; aDCF loss; speaker recognition; deep neural networks; spectral clustering; video processing

1. Introduction

A multimodal biometric verification field consists of the identification of persons by means of more than one biometric characteristics, as the use of two modalities makes the process more robust to potential problems. Typically, face and voice characteristics have been two of the preferred biometric data due to the ease of obtaining audiovisual resources to carry out the systems that perform this process. When this identification process is applied throughout a video file, and this information is kept over time, this kind of task is also known as multimodal diarization combined with identity assignment. In recent years, this field has been widely investigated due to its great interest, motivated by the fact that human perception uses not only acoustic information but also visual information to reduce speech uncertainty. Moreover, this task has been rarely addressed for uncontrolled data due to the lack of this type of datasets. However, several challenges focused on this topic have recently been developed [1–3], and a large amount of multimedia and broadcast data is also currently being produced, such as news, talk shows, debates, or series. Therefore, to develop a multimodal biometric system, different tools are required to process this data, detect the presence of people, and address the identification of who is appearing and speaking. The need to find new, efficient tools to process all the available audiovisual content has led to a wide variety of systems based on artificial intelligence algorithms, such as deep neural networks (DNN).

To perform multimodal diarization, many studies focus on the simplest method which is based on independent systems for speaker and face diarization [3,4]. Speaker diarization is a widespread task [5,6] due to its usefulness as preprocessing 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 used as a video indexing tool, and as the previous step for face verification [7,8]. However, in real-world scenarios, face images can often appear with large variations, so this kind of system has also found some problems in unconstrained videos. For these reasons, a straightforward score-level fusion is usually employed to join the information from both types of systems.

The IberSPEECH-RTVE 2020 Challenges aim to benchmark and further analyze this different kind of diarization systems. Therefore, two types of diarization evaluations are included in this challenge: speaker diarization and identity assignment (SDIA) [9], and 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 and the face identity assignment, which is obtaining more relevance in recent times. In this work, we focused on this second challenge and, especially, the characteristics of the face subsystem are remarked upon.

This paper presents the ViVoLAB system submitted to the IberSPEECH-RTVE 2020 Challenge in the MD task. This challenge focuses on the segmentation of broadcast audiovisual documents and the assignment to segments of an identity from a closed set of different faces and speakers. The face and speaker identities from this closed set are known as enrollment or target identities. For the challenge, we processed audio and video 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, both audio and video files are processed. After that, an embedding extractor is used to obtain the representations and, finally, clustering and assignment process is applied. The assignment process can be seen as a binary task that consists of comparing each face or speaker present in the audiovisual file with all the enrollment identities and determining whether it belongs to one of them or not. A simple approach employed is a cosine similarity by averaging the representations of all the enrollment files for each identity to obtain the verification scores and decide the identity. Nevertheless, these representations are extracted from DNN systems which are not trained with this objective, so instead of using only cosine similarity, complex back-ends [11,12] are often applied to improve the discriminative ability of these representations. The drawbacks of this kind of back-end are that it involves a more complex training process and, therefore, a high computational cost. Thus, to carry out the assignment process in the face subsystem of this work, a new approach based on [13] was applied to model the enrollment identities. This new approach was shown to be a promising technique to characterize each enrollment identity with a single learnable vector for the speaker verification task, but this is the first time that this technique has been applied in face verification. To train this back-end, the approximated detection cost function (aDCF) [14] was used as the objective loss. Hence, in this work, different parameter settings of this loss were explored and their effect on the errors produced by the system was studied.

This paper is organized as follows. In Section 2, we provide a description of the challenge and the dataset employed. Section 3 describes the new approach based on training face enrollment models by network optimization and the loss function used as objective for the training. Section 4 details the face diarization subsystem. The speaker diarization employed is explained in Section 5. In Section 6, the performance metric employed is detailed. Finally, Section 7 presents and discusses results, and Section 8 concludes the paper.

2. RTVE 2020 Challenge

The RTVE 2020 Challenge is part of the 2020 edition of the Albayzin evaluations [10,15]. This dataset is a collection of several broadcast TV shows in Spanish language covering different scenarios. To carry out this challenge, the database provides around 40 h of shows from Radio Televisión Española (RTVE), the Spanish public radio and television. The development subset of the RTVE2020 database contains two of the parts of the RTVE 2018 database (*dev*2 and *test* partitions) which are formed by four shows of around 6 h. Furthermore, this subset also contains a new development partition with nine shows of around 4 h. The evaluation or test set consists of fifty-four video files of around 29 h in total with timestamps of speakers and faces. Enrollment data is also provided for 161 characters with 10 images and a 20 s video of each character.

3. Face Enrollment Models

In verification tasks, a back-end is traditionally applied to compare enrollment and test embeddings and obtain the final verification scores to assign the correspondent identity. A widely used approach is the cosine similarity, where if an enrollment identity has more than one enrollment embedding, these embeddings are averaged to compare with the test embedding. However, we have shown in [13] for the speaker verification task that a better solution to perform this process consists of training an enrollment model for each enrollment identity. Therefore, in this work, we applied this approach for the face verification task where we trained a model for each of the face enrollment identities. The loss function optimized and the process to carry out the training of these models with this loss function are explained in detail below.

3.1. Approximated Detection Cost Function (aDCF) Loss

Most DNN systems are trained to generate representations using traditional loss functions as the objective loss for training. However, this strategy has a main drawback, as traditional loss functions are not oriented to the goal task. For this reason, different alternatives have been presented in the literature to design loss functions focused on the final evaluation metrics to train the DNN systems such as the approximated area under the ROC curve (aAUC) [16,17], the partial and multiclass AUC loss (pAUC) [18–20], and the approximated detection cost function (aDCF) [14] which was used for this work. This aDCF is inspired by the DCF verification metric [21]. The use of this differentiable version of the DCF metric as objective loss function allows training DNN systems aimed at minimizing one of the main metrics employed in verification tasks. In addition, this function is based on measuring the two types of decision errors produced by verification systems using a threshold. These two errors are known as false alarms and misses which are also part of the diarization error rate (DER) used to evaluate diarization systems. The former errors occur when an identity is incorrectly assigned, while the latter refer to a correct identity not being detected by the system. Therefore, we seek to minimize the average number of times false alarms (P_{fa}) and misses (P_{miss}) occur, which can be approximated as

$$\hat{P}_{fa}(\theta,\Omega) = \frac{\sum_{y_i \in y_{non}} \sigma_\alpha(s_\theta(x_i,y_i) - \Omega)}{N_{non}}, \tag{1}$$

$$\hat{P}_{miss}(\theta,\Omega) = \frac{\sum_{y_i \in y_{tar}} \sigma_\alpha(\Omega - s_\theta(x_i,y_i))}{N_{tar}}, \tag{2}$$

where x_i is the input sample, y_i is the class label, $s_\theta(x_i,y_i)$ is the score obtained from the last layer which is defined as a cosine distance layer, and $\sigma_\alpha(\cdot)$ is the sigmoid function, expressed as follows:

$$\sigma_\alpha(s) = \frac{1}{1 + exp(-\alpha \cdot s)}, \tag{3}$$

where α is an adjustable parameter. The use of this sigmoid function allows making the original expressions of P_{fa} and P_{miss} differentiable and enables the backpropagation of gradients. Hence, the aDCF loss function to minimize is composed of a weighted sum of these approximated expressions, defined by

$$aDCF(\theta, \Omega) = \gamma \cdot \hat{P}_{fa}(\theta, \Omega) + \beta \cdot \hat{P}_{miss}(\theta, \Omega), \tag{4}$$

where γ and β are configurable parameters to provide more cost relevance to one of the terms over the other.

3.2. Training Process of Enrollment Models

Motivated by the demonstrated effectiveness in training DNNs using aDCF loss function, in [13], we developed a straightforward and powerful back-end approach based on a network optimization with this loss function. This approach tries to mimic the target/nontarget process performed in verification tasks. In addition, this back-end takes advantage of the data learned during a previous training step of a general embedding network. Thus, this approach avoids the need for a careful selection of input data to train the models as required by other complex back-ends such as triplet neural network with triplet loss [11] or triplet neural network combined with aAUC [16,17].

Figure 1 shows the process to carry out this training, where a learnable vector is obtained to represent each enrollment identity. This process is based on comparing the positive or target examples with themselves (s_{tar}), and also with the negative or nontarget examples (s_{nontar}), using the aDCF loss function as training objective loss. To optimize aDCF loss, the scores used are obtained with cosine similarity as

$$s_{\theta}(x_i) = \frac{x_i \cdot w^T}{\|x_i\| \cdot \|w^T\|}, \tag{5}$$

where $\|x_i\|$ is the normalized input to the enrollment model, and $\|w^T\|$ is the normalized layer parameters of the embedding obtained from the enrollment utterance.

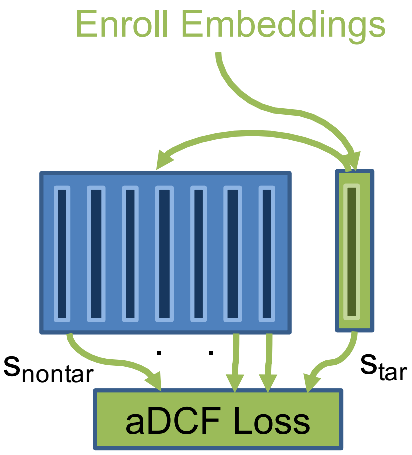


Figure 1. Training face enrollment models using target and nontarget embeddings for each enroll or target identity.

The philosophy of the approach followed in this work has been the same as the original approach used for speaker verification, but in this work, there are several differences. First, in our previous work, the final verification scores were obtained directly in the last step of the training process by comparing all target and nontarget samples with the trained vector, whereas in this work, the learnable vector is stored as an enrollment model to be used in the final step of the face subsystem to assign the identity to each segment. On the other hand, the nontarget examples employed in this system are directly the embeddings extracted from the pretrained model. These nontarget examples belong to identities different to the enrollment identities, so we used them to train the enrollment models instead of using the weight matrix of the last layer of the trained neural network to obtain them as we did in [13]. Hence, in this work, the process for training the enrollment model for each identity is based on the following steps:

1. The target and nontarget embeddings extracted from the pretrained model are employed as positive and negative examples.
2. Each enrollment model is trained using the aDCF loss function with all nontarget embeddings and only the target embeddings of the corresponding enrollment identity.
3. The trained models are stored to use them in the assignment process.

4. Face Subsystem

In this section, we present 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 Figure 2.

4.1. Video Processing

The video processing step used to develop this face subsystem consists of three blocks: frame extraction, face detection, and change shot detection. In the following, we will detail all of them.

4.1.1. Frame Extraction

As the first step, the video is processed to extract five frames per second using the $ffmpeg$ tool (https://www.ffmpeg.org/ (accessed on 23 December 2021)). We decided to use five frames per second as this number of frames allows us a high precision to determine the limits of the characters appearance. Therefore, frames are extracted using a constant rate where one frame is obtained every 200 ms.

4.1.2. Face Detection

Another fundamental step in this process is the face detection because failures in this step could be crucial for the 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) [22]. In this part, we employ this implemented system as it is a proven and effective method for face detection, which is necessary to perform before continuing with the rest of the face verification pipeline. Furthermore, using this detector, we can store the bounding boxes created by the algorithm that correspond to the coordinates where a face is detected. This information is then employed in the tracking and identity assignment process.

4.1.3. 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 characteristics of the content and by constant changes of shots and scenes. Thus, to aid the tracking and clustering step, a change scene detection tool (https://www.pyscenedetect.readthedocs.io/en/latest/ (accessed on 23 December 2021)) is employed, as this tool effectively detects these changes using threshold-based detection mode. This detector finds areas where the difference between two subsequent frames exceeds a threshold value.

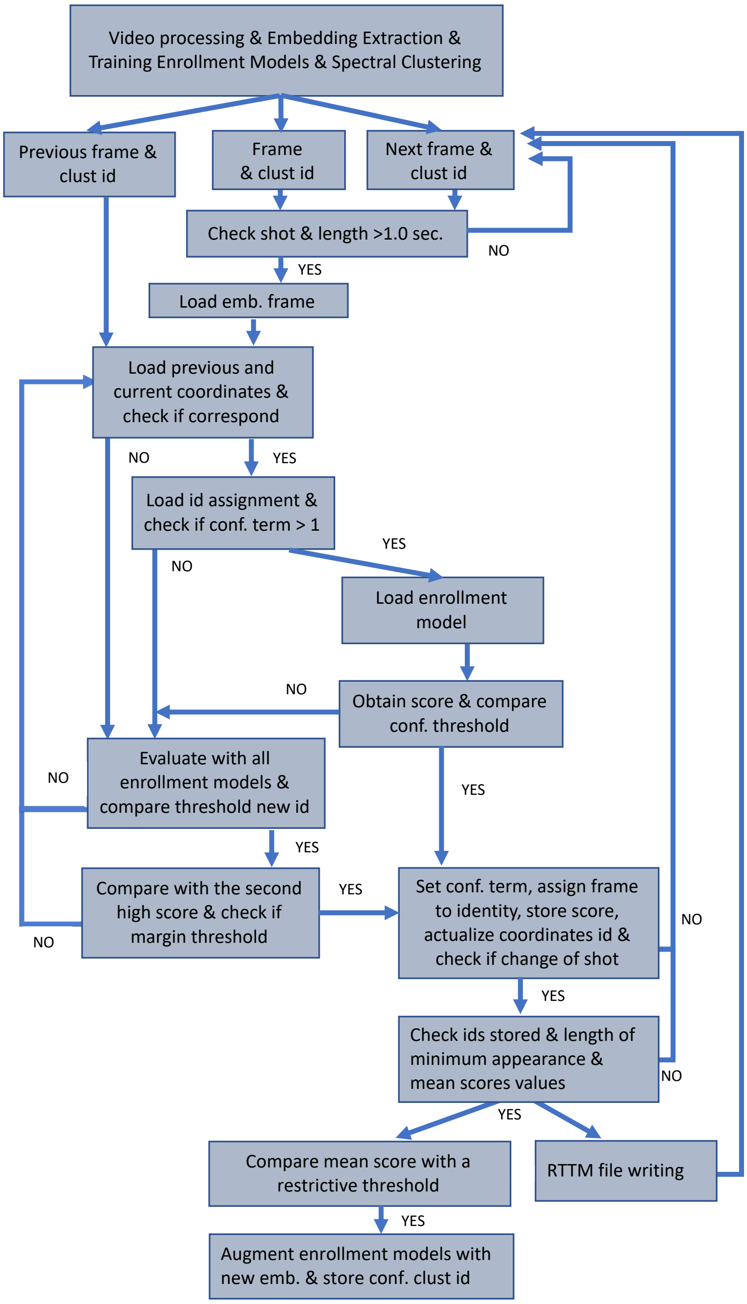

Video processing & Embedding Extraction & Training Enrollment Models & Spectral Clustering
Previous frame & clust id
Frame & clust id
Next frame & clust id
Check shot & length >1.0 sec.
NO
YES
Load emb. frame
Load previous and current coordinates & check if correspond
NO
YES
Load id assignment & check if conf. term > 1
YES
NO
Load enrollment model
NO
Obtain score & compare conf. threshold
NO
Evaluate with all enrollment models & compare threshold new id
YES
YES
Compare with the second high score & check if margin threshold
YES
NO
Set conf. term, assign frame to identity, store score, actualize coordinates id & check if change of shot
NO
YES
Check ids stored & length of minimum appearance & mean scores values
NO
YES
Compare mean score with a restrictive threshold
RTTM file writing
YES
Augment enrollment models with new emb. & store conf. clust id

Figure 2. Block diagram of face system.

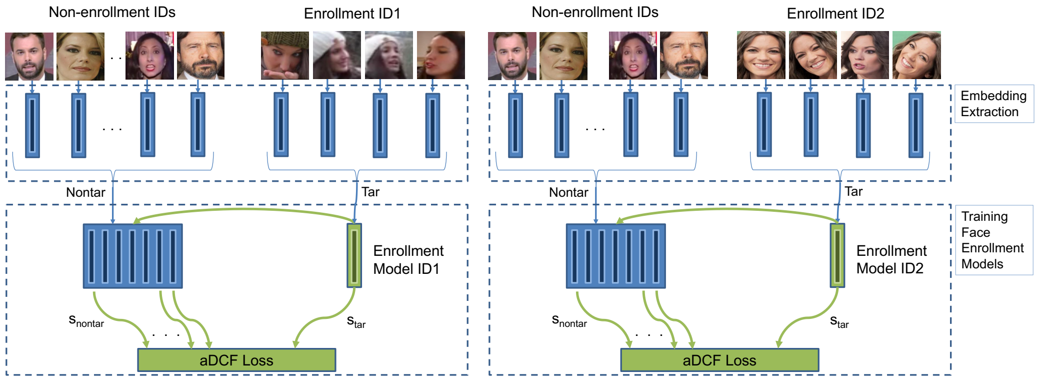
4.2. Embedding Extraction

Once the video processing step is complete, we process the face images using the bounding boxes, and apply mean and variance normalization. Then, as the images were processed with the information given by the face detector, the resulting images have centered faces. Therefore, a center crop can be applied to resize the images to 160 × 160 pixels. After that, the processed images are passed through a trained model to obtain embedding representations. The indicated center crop is necessary because the model employed was trained using images with this format. In this system, as a face extractor, we employed a pretrained convolutional neural network (CNN) with more than one hundred layers [11]. This network was trained for a classification task on the CASIA-WebFace dataset [23], but the embeddings extracted from it have been proved previously in a verification task to check their discriminative ability with state-of-the-art results on Labeled Faces in the Wild

(LFW) [24,25]. For this reason, we decided to use these embeddings of 128 dimensions to extract the representations of the enrollment and test files of this challenge.

4.3. Training Face Enrollment Models

As we explained in Section 3, in this work, we applied the approach based on enrollment models [13] as a back-end for the face verification task where we trained a model for each one of the 161 enrollment identities. To train these models, we used the embeddings of the enrollment images, and the video files of the development and test sets of the IberSPEECH-RTVE 2020 Challenge [10] as positive or target examples, while the enrollment files of the development and test sets of the IberSPEECH-RTVE 2018 Challenge [15] were used as negative or nontarget examples. Therefore, once these embeddings were extracted, we trained each face enrollment model with them using the aDCF loss function. Figure 3 shows two examples of the steps presented in Section 3.2 of the process of extracting embeddings and training an enrollment model for each identity using the aforementioned data. Moreover, the impact of the amount of nontarget data employed to train these models will be discussed in the experimental section.



(**a**) Model ID1 (**b**) Model ID2

Figure 3. (**a**) Left: Example of Embedding Extraction and Training Enrollment Model ID1, where the dashed line indicates the two steps of the process. (**b**) Right: Example of Embedding Extraction and Training Enrollment Model ID2, where the dashed line indicates the two steps of the process.

4.4. Clustering

As a source of complementary information, the face embeddings from the test videos are used to perform a spectral clustering technique [26] that attempts to find strongly connected segments. This technique provides an initial cluster assignment to group the frames of the video sequence. In this work, we employed this clustering combined with the use of coordinates to improve the whole tracking process.

4.5. Tracking and Identity Assignment Scoring

Once all the above information was obtained, we developed an algorithm to carry out the tracking and identity assignment process, which is depicted in Figure 2 and follows a similar philosophy to the one developed in [27]. In this algorithm, the tracking process was developed by shot, so a change of shot restarts the tracking. Therefore, while the shot 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 to perform the tracking process. When a relationship exists between both frames 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 assigned identity is correct or not. However, when there is no relationship between the coordinates of the current and previous frame, or the confidence term is low, the frame embedding is compared to all enrollment models to obtain a score and determine whether it is a new identity to assign. Once the identity assignment is performed on the current frame, the score is stored, the coordinates are updated, and the algorithm checks whether the shot changes.

Tracking is carried out with the above steps, but the identity assignment process performed is only an initial assignment. When a shot change is detected, the system checks the identities and scores stored in the shot to remove inconsistent segment assignments. After that, the final segments with their identity assignments are written to 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 detected shots.

5. Speaker Subsystem

This section describes the speaker subsystem, which consists of similar blocks to the face subsystem, such as audio processing, embedding extraction, clustering, and identity assignment scoring, but using different approaches in each one. The different parts of just this subsystem are explained in more detail in [28], although, as part of the multimodal diarization system, the following sections present the main ideas for creating this subsystem.

5.1. Audio Processing

In this subsystem, the audio processing step is also composed of three blocks: a front-end, speech activity detection, and speaker change point detection. Next, we briefly explain each of them.

5.1.1. Front-End and Speech Activity Detection

The first block of this subsystem is a front-end to obtain the MFCC features [29]. For a given audio, a stream of 32 coefficient feature vectors is estimated according to a 25 ms window with a 15 ms overlap. No derivatives are considered. Simultaneously, speech activity detection (SAD) labels are estimated each as 10 ms [30]. Our approach for SAD is based on a deep learning solution that is an evolution derived from our previous SAD systems [31]. We use a convolutional recurrent neural network (CRNN) consisting of three blocks of 2D convolutional followed by three BiLSTM layers [32]. 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 fed to the neural network. Cepstral mean and variance normalization (CMVN) [33] normalization is applied.

The CRNN is trained on a combination of different broadcast datasets. Specifically, we include data from the Albayzín 2010 dataset (train and eval), Albayzín 2018 dataset (dev2 and eval), and a selection of data from the first MGB Challenge (train, dev.longitudinal, and task3 eval). Furthermore, audios are augmented with a variety of noises that can be usually found in broadcast emissions.

5.1.2. Speaker Change Point Detection

Feature vectors and SAD labels obtained are fed into the speaker change point detection (SCPD) block which is dedicated to infer the speaker turn boundaries. The differential form of Bayesian information criterion (ΔBIC) [34] was used. This estimation works in terms of a 6 s sliding window, in which we assume there is, at most, a speaker turn boundary. Each involved speaker in the analysis is modeled by means of a full-covariance Gaussian distribution. In addition, the SAD labels delimit the parts of the audio where the analysis is performed. In the given data, the described configuration provides segments of approximately 3 s length on average.

5.2. Embedding Extraction

Once the audio processing is completed, each one of the identified segments will be converted into a compact representation or embedding. For this purpose, we opted for an evolution of x-vectors [35] considering an extended version [36] of the time delay neural network (TDNN) architecture [37]. The modification is the inclusion of multi-head self-attention [38] in the pooling layer. This network, trained on VoxCeleb [39] and VoxCeleb2 [40], provides embeddings of dimension 512. These embeddings undergo centering and LDA whitening (reducing dimension to 200), both trained with MGB [41] as well as the Albayzín training subset, and finally length normalization [42]. These embeddings will be referred to as Φ. A similar extraction pipeline working offline is in charge of the enrollment audios. The enrollment embeddings will be named Φ_{enroll}.

5.3. Clustering

The obtained embeddings are modeled in a generative manner according to [43], where a tree-based probabilistic linear discriminant analysis (PLDA) clustering is proposed. This approach exploits the higher acoustic similarity of temporally close embeddings by sequentially assigning these representations to the available clusters at each time. These clusters are managed by the algorithm at the same time. This concept is boosted by [44], which helps to find the best possible sequence of decisions. The considered model has 100-dimension speaker subspace and it is trained with both MGB and Albayzín training subset.

5.4. Identity Assignment Scoring

The considered identity assignment block follows a state-of-the-art speaker recognition PLDA backend followed by score normalization and calibration stages. By means of the PLDA model, we estimate the score s_{ij}, which represents the likelihood that the diarization cluster j shares the same speaker identity as the enrollment speaker i. Then, these scores are normalized according to adaptive S-norm [45], using MGB as extra cohort. Finally, normalized scores are calibrated according to a threshold ϵ. Whenever the score s_{ij} overcomes the threshold, we consider that the cluster j contains audio from the enrolled person i, being different otherwise. Threshold ϵ is adjusted by assignment error rate (AER) minimization according to a calibration subset Φ_{calib} and the enrollment embeddings Φ_{enroll} as follows:

$$\epsilon = \arg\min_{\epsilon}(\text{AER}(\Phi_{\text{calib}}, \Phi_{\text{enroll}}, \epsilon)) \tag{6}$$

where Φ_{calib} and Φ_{enroll} represent the set of embeddings from calibration as well as the enrollment speakers.

Final AER labels are obtained according to these normalized and calibrated scores. The audio from cluster j is assigned to the ith enrolled identity with highest score if s_{ij} overcomes the calibration threshold. If s_{ij} is below the threshold for any enrolled identity i, the cluster is assigned to the generic unknown identity. Mathematically, the assigned identity (θ_j) for a subset of embeddings j with respect to the enrolled identity i is

$$\theta_j = \begin{cases} \arg\max_i(s_{ij}|s_{ij} > \epsilon) & \text{if } \exists i|s_{ij} > \epsilon \\ \text{Unknown} & \text{if } \forall i, s_{ij} < \epsilon \end{cases} \tag{7}$$

where s_{ij} stands for the normalized PLDA log-likelihood ratio score between the embedding j and the enrolled identity i. By means of this decision-making, we do not exclude the possibility of assigning multiple diarization clusters to the same identity. This design choice is taken to allow the fix of some diarization errors.

6. Performance Metrics

To evaluate the systems developed in this work, the metric used was diarization error rate (DER). DER is usually the reference metric employed in the diarization task, but in this case, DER is obtained slightly differently than the original metric as it also takes into account the measurement of the identity assignment errors. To better analyze the results obtained with the DER metric, this metric can be decomposed in the three terms of error:

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

7. Results

In this work, several experiments were carried out to show the effect of different aspects on the face subsystem and the overall performance of both subsystems. First, we compared the use of a cosine similarity metric directly on the embeddings extracted from the pretrained model ($AverageEmbedding$) to obtain the closest identity in each instance with the training face enrollment models approach ($EnrollmentModels$) for the identity assignment process. Moreover, in this first set of experiments, the relevance of employing more nontarget data to train the enrollment models was also analyzed. After that, we analyzed the behaviour of the system when different values of the aDCF loss function parameters were employed. To conclude, a summary of the best results of the face subsystem in combination with the results of the speaker subsystem is presented.

7.1. Analysis of Training Enrollment Models for Face Subsystem

In this section, we analyze the performance of the system when enrollment models are trained and used for the identity assignment process or a cosine similarity is directly employed to compare the average of all enrollment embeddings with each frame of the video file. Furthermore, the effect of adding more nontarget data to train the enrollment models is also checked.

Table 1 shows DER% results on the test set for the face subsystem with the different back-end approaches. As we can observe, the training of face enrollment models to characterize each enrollment identity achieves a large improvement over comparing each segment directly against the average of the enrollment embeddings. Note that whether the enrollment models are trained with more nontarget examples, the variability that the learnable vectors have to model is higher, so these models learn to represent each identity better and the DER% result obtained is lower.

Table 1. Experimental results on RTVE 2020 Multimodal Diarization test set, showing DER%. These results were obtained to compare the back-end approach proposed and the cosine baseline. The best result is marked in bold.

Back-End	Nontarget Examples	DER%
Average Embedding	–	80.16%
Enrollment Models	57	61.79%
	3302	**56.86%**

7.2. Effect of aDCF Parameters γ, β for Training Face Enrollment Models

A second set of experiments was performed to observe the effect of training with different aDCF loss parameters. In Table 2 and Figure 4, we observe that the system performance improves by only adjusting the aDCF parameters without modifying the threshold values or the tracking and identity assignment algorithm for the reference number of training epochs. As reference number of epochs, we considered 800 epochs as it is the number initially used in the previous set of experiments. In view of this result, we explored, in depth, the behavior of the training enrollment models for the different configurations of parameters and number of epochs. As a result of this sweep, we obtained that the original parameter configuration, $\gamma = 0.75$ and $\beta = 0.25$, has still room for improvement, and training for 1200 epochs achieves the best result without modifying any other parameter. Moreover, we note that when giving a higher cost relevance (β) to the probability of misses during training with aDCF loss function, regardless of the number of epochs, the results are worse in all situations.

Table 2. Experimental results of RTVE 2020 Multimodal Diarization test set, showing DER%. These results were obtained by sweeping of the parameters of aDCF loss function and by different number of training epochs. The best results are marked in bold.

γ	β	600 Epochs	800 Epochs	1000 Epochs	1200 Epochs	1400 Epochs
0.75	0.25	62.56%	56.86%	**55.32**%	**54.07**%	**55.91**%
0.50	0.50	**55.92%**	**55.89%**	56.02%	56.47%	57.16%
0.25	0.75	58.30%	59.29%	59.58%	60.01%	60.02%

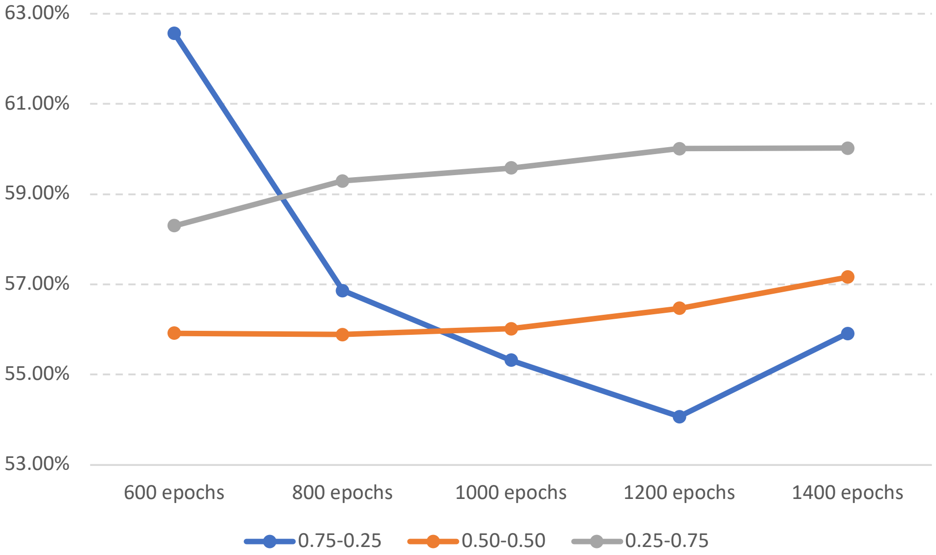


Figure 4. Evolution of DER% results as a function of the different parameter configurations.

In addition to the above results, we analyzed the behavior of the three types of errors that compose the DER% metric when the different possible system configurations are used for the same number of training epochs. As we can see in Table 3 and Figure 5, giving a higher cost relevance to the probability of false alarms (γ) in the aDCF loss results in a lower number of false alarms in the decomposition of the DER%, while the number of misses is higher than in the other two configurations. On the other hand, the same trend can be seen in reverse when the relevance term (β) is higher for the probability of misses.

Table 3. Experimental results of RTVE 2020 Multimodal Diarization test set, showing DER% and decomposition of DER% results in miss (MISS), false alarm (FA), and identity (ID) errors. These results were obtained by sweeping the parameters of the aDCF loss function. The best results are marked in bold.

γ	β	MISS	FA	ID	DER
0.75	0.25	32.60%	**12.80%**	**8.70%**	**54.07%**
0.50	0.50	30.70%	16.00%	9.80%	56.47%
0.25	0.75	**28.30%**	19.80%	12.00%	60.01%

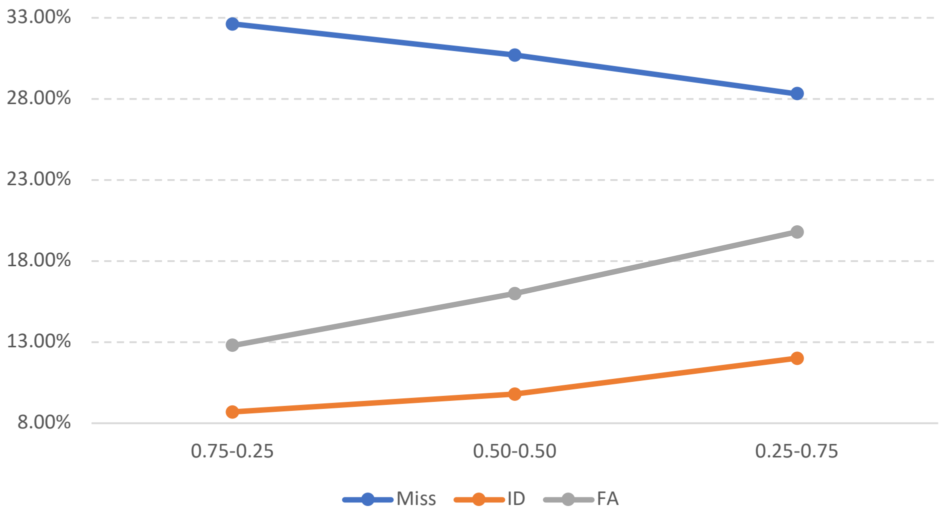


Figure 5. Evolution of the different types of errors (MISS, FA, ID) for each different γ, β parameter configuration.

7.3. Summary of Face and Speaker Results

In this section, we collect the best result for the face and speaker subsystems, and we also divide the results into development and test set to better observe the difference in behavior of both subsystems. Moreover, our reference results obtained for the challenge [46] are included to better reflect the improvement achieved. Thus, the results of the other two participating groups [47,48] that are publicly available (available online: http://catedrartve.unizar.es/albayzin2020results.html (accessed on 23 December 2021)) are been introduced in reference to the difficulty of this multimodal diarization challenge.

Table 4 shows the DER results obtained in the development and test set for the face and speaker modalities. In addition to the separate results, we show the average result of the face and speaker diarization errors ($FACE + SPEAKER$). These results indicate a great mismatch between development and test results. We analyzed what type of video files composed both subsets and the length of these files, and we found that the development files are shorter and more similar than test files. Thus, we can see that the face and speaker subsystems perform better in the development files which are shorter videos, so the tracking process is easier to follow. Nevertheless, in the face subsystem, we observe that this difference is smaller than in the speaker subsystem.

Table 4. Experimental results on RTVE 2020 Multimodal Diarization development and test sets, showing DER%. These DER% values were the result of the improvements introduced in this work, and the reference results for both modalities are also presented.

Subset	Modality	DER%	DER% Ours [46]	DER% [47]	DER% [48]
DEV	FACE	51.26%	51.66%	–	–
	SPEAKER	37.45%	47.90%	–	–
	FACE+SPEAKER	44.36%	49.78%	–	–
TEST	FACE	54.07%	61.79%	44.55%	67.31%
	SPEAKER	60.34%	72.63%	61.61%	131.59%
	FACE+SPEAKER	57.20%	67.21%	53.08%	99.45%

To better analyze these results, Table 5 presents a decomposition of the DER metric into the three terms of error. Focusing on the face modality errors, in the case of the development subset, we observe that the main cause of error is the probability of misses, which indicates that a large number of segments of the target identities have not been detected. Therefore, this effect can be motivated by using a threshold value that is too restrictive, while in the test subset, the misses term decreases, and especially relevant is the increase in false alarm errors as this illustrates the problems in discarding segments of nontarget 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, as false alarms are much larger than misses in the test subset of data. Note that it is also related to the chosen threshold; however, in this case, the threshold is lower, so the target segments are mostly detected, but as a result, a high number of enrollment identities are assigned to segments of unknown identity.

Table 5. 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	41.70%	5.10%	4.50%
	TEST	32.60%	12.80%	8.70%
SPEAKER	DEV	26.10%	9.60%	1.75%
	TEST	8.70%	39.00%	12.64%

8. Conclusions

This paper presents the ViVoLAB submission to the IberSPEECH-RTVE 2020 Multimodal Diarization Challenge. In this work, we developed two monomodal subsystems to separately address face and speaker diarization. Each system is based on state-of-the-art DNN approaches. In addition, we introduced a new back-end approach for the face subsystem. This approach consists of training a learnable vector with the aDCF loss function to represent each face enrollment identity. Using these enrollment models for the identity assignment process instead of just the cosine similarity, the results have achieved a relevant improvement over the average embedding directly and the application of cosine similarity. 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 carried out on the fusion of both systems, which could improve the final results, especially by disambiguating the identification process. The high DER values for misses and false alarms in the face and speaker subsystem, respectively, should be addressed by that fusion.

Author Contributions: Conceptualization, V.M. and A.M.; Investigation, V.M. and A.M.; Methodology, V.M. and A.M.; Software, V.M., I.V., P.G. and A.M.; Supervision, A.M., A.O. and E.L.; Writing—original draft, V.M.; Writing—review and editing, A.M., A.O. and E.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant 101007666; in part by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU"/PRTR under Grant PDC2021-120846-C41, by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the grant PRE2018-083312, and by the Government of Aragón (Grant Group T36_20R).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Poignant, J.; Bredin, H.; Barras, C. Multimodal Person Discovery in Broadcast tv at Mediaeval 2015. MediaEval 2015 Working Notes Proceedings. 2015. Available online: CEUR-WS.org (accessed on 23 December 2021).
2. Bredin, H.; Barras, C.; Guinaudeau, C. Multimodal Person Discovery in Broadcast TV at MediaEval 2016. MediaEval 2016 Working Notes Proceedings. 2016. Available online: CEUR-WS.org (accessed on 23 December 2021).
3. Sadjadi, O.; Greenberg, C.; Singer, E.; Reynolds, D.; Mason, L.; Hernandez-Cordero, J. The 2019 NIST Audio-Visual Speaker Recognition Evaluation. In Proceedings of the Odyssey 2020 The Speaker and Language Recognition Workshop, Tokyo, Japan, 1–5 November 2020; pp. 259–265.
4. Das, R.K.; Tao, R.; Yang, J.; Rao, W.; Yu, C.; Li, H. HLT-NUS Submission for NIST 2019 Multimedia Speaker Recognition Evaluation. *arXiv* **2020**, arXiv:2010.03905.

5. Garcia-Romero, D.; Snyder, D.; Sell, G.; Povey, D.; McCree, A. Speaker diarization using deep neural network embeddings. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4930–4934.
6. Viñals, I.; Gimeno, P.; Ortega, A.; Miguel, A.; Lleida, E. In-domain Adaptation Solutions for the RTVE 2018 Diarization Challenge. In Proceedings of the Iberspeech 2018, Barcelona, Spain, 21–23 November 2018; pp. 220–223.
7. Khoury, E.; Gay, P.; Odobez, J.M. Fusing matching and biometric similarity measures for face diarization in video. In Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, Dallas, Texas, USA, 16–19 April 2013; pp. 97–104.
8. Le, N.; Heili, A.; Wu, D.; Odobez, J.M. Efficient and Accurate Tracking for Face Diarization via Periodical Detection. In Proceedings of the International Conference on Pattern Recognition, Cancun, Mexico, 4–8 December 2016.
9. Ortega, A.; Miguel, A.; Lleida, E.; Bazán, V.; Pérez, C.; Gómez, M.; de Prada, A. Albayzin evaluation: IberSPEECH-RTVE 2020 Speaker Diarization and Identity Assignment. Available online: http://catedrartve.unizar.es/reto2020/EvalPlan-SD-2020-v1.pdf (accessed on 23 December 2021).
10. Lleida, E.; Ortega, A.; Miguel, A.; Bazán, V.; Pérez, C.; Gómez, M.; de Prada, A. Albayzin evaluation: IberSPEECH-RTVE 2020 Multimodal Diarization and Scene Description Challenge. Available online: http://catedrartve.unizar.es/reto2018/EvalPlan-Multimodal-v1.3.pdf (accessed on 23 December 2021).
11. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A unified embedding for face recognition and clustering . In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823. [CrossRef]
12. Jung, J.; Heo, H.; Kim, J.; Shim, H.; Yu, H. RawNet: Advanced End-to-End Deep Neural Network Using Raw Waveforms for Text-Independent Speaker Verification. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1268–1272.
13. Mingote, V.; Miguel, A.; Ortega, A.; Lleida, E. Training Speaker Enrollment Models by Network Optimization. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 3810–3814.
14. Mingote, V.; Miguel, A.; Ribas, D.; Ortega, A.; Lleida, E. Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 2903–2907.
15. Lleida, E.; Ortega, A.; Miguel, A.; Bazán-Gil, V.; Pérez, C.; Gómez, M.; de Prada, A. Albayzin 2018 evaluation: The iberspeech-RTVE challenge on speech technologies for spanish broadcast media. *Appl. Sci.* **2019**, *9*, 5412. [CrossRef]
16. Mingote, V.; Miguel, A.; Ortega, A.; Lleida, E. Optimization of the area under the ROC curve using neural network supervectors for text-dependent speaker verification. *Comput. Speech Lang.* **2020**, *63*, 101078. [CrossRef]
17. Mingote, V.; Castan, D.; McLaren, M.; Nandwana, M.K.; Ortega, A.; Lleida, E.; Miguel, A. Language Recognition Using Triplet Neural Networks. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 4025–4029.
18. Bai, Z.; Zhang, X.; Chen, J. Partial AUC Optimization Based Deep Speaker Embeddings with Class-Center Learning for Text-Independent Speaker Verification. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6819–6823.
19. Gimeno, P.; Mingote, V.; Ortega, A.; Miguel, A.; Lleida, E. Partial AUC Optimisation Using Recurrent Neural Networks for Music Detection with Limited Training Data. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 3067–3071.
20. Gimeno, P.; Mingote, V.; Ortega, A.; Miguel, A.; Lleida, E. Generalising AUC Optimisation to Multiclass Classification for Audio Segmentation with Limited Training Data. *IEEE Signal Process. Lett.* **2021**, *28*, 1135–1139. [CrossRef]
21. Martin, A.; Przybocki, M. The NIST 1999 speaker recognition evaluation—An overview. *Digit. Signal Process.* **2000**, *10*, 1–18. [CrossRef]
22. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
23. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning face representation from scratch. *arXiv* **2014**, arXiv:1411.7923.
24. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In Proceedings of the Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition, Marseille, France, 12–18 October 2008.
25. Huang, G.; Mattar, M.; Lee, H.; Learned-Miller, E.G. Learning to align from scratch. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 764–772.
26. Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905.
27. Ramos-Muguerza, E.; Docío-Fernández, L.; Alba-Castro, J.L. The GTM-UVIGO System for Audiovisual Diarization. In Proceedings of the Iberspeech 2018, Barcelona, Spain, 21–23 November 2018; pp. 204–207.
28. Viñals, I.; Ortega, A.; Miguel, A.; Lleida, E. The Domain Mismatch Problem in the Broadcast Speaker Attribution Task. *Appl. Sci.* **2021**, *11*, 8521. [CrossRef]
29. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [CrossRef]

30. Gimeno, P.; Ribas, D.; Ortega, A.; Miguel, A.; Lleida, E. Convolutional recurrent neural networks for speech activity detection in naturalistic audio from apollo missions. In Proceedings of the IberSPEECH 2021, Valladolid, Spain, 24–25 March 2021; pp. 26–30.
31. Viñals, I.; Gimeno, P.; Ortega, A.; Miguel, A.; Lleida, E. Estimation of the Number of Speakers with Variational Bayesian PLDA in the DIHARD Diarization Challenge. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 2803–2807.
32. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
33. Alam, M.J.; Ouellet, P.; Kenny, P.; O'Shaughnessy, D. Comparative evaluation of feature normalization techniques for speaker verification. In Proceedings of the International Conference on Nonlinear Speech Processing, Las Palmas de Grancanaria, Spain, 7–9 November 2011; pp. 246–253.
34. Chen, S.; Gopalakrishnan, P. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, USA, 8–11 February 1998; Volume 8, pp. 127–132.
35. Snyder, D.; Ghahremani, P.; Povey, D.; Garcia-Romero, D.; Carmiel, Y.; Khudanpur, S. Deep neural network-based speaker embeddings for end-to-end speaker verification. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016; pp. 165–170.
36. Villalba, J.; Chen, N.; Snyder, D.; Garcia-Romero, D.; McCree, A.; Sell, G.; Borgstrom, J.; Richardson, F.; Shon, S.; Grondin, F.; et al. State-of-the-Art Speaker Recognition for Telephone and Video Speech: The JHU-MIT Submission for NIST SRE18. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1488–1492.
37. Waibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K.; Lang, K.J. Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 328–339. [CrossRef]
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
39. Nagrani, A.; Chung, J.S.; Zisserman, A. VoxCeleb: A Large-Scale Speaker Identification Dataset. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 2616–2620.
40. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxCeleb2: Deep Speaker Recognition. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 1086–1090.
41. Bell, P.; Gales, M.J.; Hain, T.; Kilgour, J.; Lanchantin, P.; Liu, X.; McParland, A.; Renals, S.; Saz, O.; Wester, M.; et al. The MGB challenge: Evaluating multi-genre broadcast media recognition. In Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 687–693.
42. Garcia-Romero, D.; Espy-Wilson, C.Y. Analysis of i-vector length normalization in speaker recognition systems. In Proceedings of the Twelfth Annual Conference of the International Speech Communication Association, Florence, Italy, 27–31 August 2011.
43. Viñals, I.; Gimeno, P.; Ortega, A.; Miguel, A.; Lleida, E. ViVoLAB Speaker Diarization System for the DIHARD 2019 Challenge. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 988–992.
44. Viterbi, A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* **1967**, *13*, 260–269. [CrossRef]
45. Brümmer, N.; Strasheim, A. Agnitio's speaker recognition system for evalita 2009. In Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, Reggio Emilia, Italy, 9–12 December 2009.
46. Mingote, V.; Vinals, I.; Gimeno, P.; Miguel, A.; Ortega, A.; Lleida, E. ViVoLAB Multimodal Diarization System for RTVE 2020 Challenge. In Proceedings of the IberSPEECH 2021, Valladolid, Spain, 24–25 March 2021; pp. 76–80.
47. Porta-Lorenzo, M.; Alba-Castro, J.L.; Docío-Fernández, L. The GTM-UVIGO System for Audiovisual Diarization 2020. In Proceedings of the IberSPEECH 2021, Valladolid, Spain, 24–25 March 2021; pp. 81–85.
48. Luna-Jiménez, C.; Kleinlein, R.; Fernández-Martınez, F.; Manuel, J.; Pardo-Munoz, J.M.M.F. GTH-UPM System for Albayzin Multimodal Diarization Challenge 2020. In Proceedings of the IberSPEECH 2021, Valladolid, Spain, 24–25 March 2021; pp. 71–75.