An End-to-End Speaker Diarization Service for improving Multimedia Content Access

DAVID MARTÍN-GUTIÉRREZ[1*], GUSTAVO HERNÁNDEZ-PEÑALOZA[2*], JOSE MANUEL MENÉNDEZ[3*], AND FEDERICO ÁLVAREZ[4*]

[*] *Visual Telecommunication Applications Group. Signals, Systems and Radio communications Department, SSR, Universidad Politécnica de Madrid*
[1] *dmz@gatv.ssr.upm.es*
[2] *ghp@gatv.ssr.upm.es*
[3] *jmm@gatv.ssr.upm.es*
[4] *fag@gatv.ssr.upm.es*

The continuous growth of Multimedia content is fostering several applied research and developments to make the content accessible to people with visual and/or hearing impairments. Broadcasters and professional content produces are aware of these limitations and they have both resources and knowledge to fulfil such needs. However, many other individuals may lack any or both of these requirements when generating and publishing multimedia content throughout the Internet. Consequently, automatic accessibility services such as automatic subtitle generation, or character detection and recognition are needed to support them. This work presents an End-to-End (E2E) Speaker Diarization architecture for automatic character recognition via Convolutional Neural Networks and Embedding Matching procedures.

1. INTRODUCTION

Speaker Diarization systems (SDS) have been widely studied and investigated to solve the problem of "who spoke when". In particular, most of the existing studies are based on the following main stages as authors remarked in [1]:

i A speech segmentation module which attempts to separate the speech parts from the non-speech ones.

ii A feature extractor stage to collect representative information from the different speakers as a set of vectors or embeddings.

iii A clustering method to both extract the number of clusters and to classify the aforementioned embeddings based on distance metrics.

iv A re-segmentation module that seeks to reinforce the segmentation of the speech parts based on the clustering results.

Moreover, many authors have proposed novel procedures to automatically extract relevant features via Convolutional Neural Networks (CNN's) to promote the acquisition of the speaker embeddings such as [2], [3], [4]. Moreover, regarding the clustering phase, most of the well-known methods are based on unsupervised techniques such as Gaussian Mixtures, Spectral clustering, or Agglomerative hierarchical clustering.

However, the aforementioned studies have some limitations that may corrupt the results in real scenarios where ambient factors such as background noise or music, overlapped sound events among others, can provoke some repercussions during the whole process. On the other hand, many of the features collected to distinguish the speakers are based on statistics or any other global features which can be improved via deep learning techniques. Moreover, the clustering method requires to set up some crucial parameters that are unknown a priori, such as the number of clusters or the minimum distance between samples in a cluster.

Furthermore, authors in [1] suggests a supervised approached that outperforms previous work in this area. This work is focused on supervised learning techniques as well to assess the identification of speakers in a certain utterance. Additionally, using speaking embeddings, this investigation proposed an architecture that can be used in many applications of speaker diarization by proposing a neural network with the capability of generalizing latent vectors for speakers.

2. END-TO-END ARCHITECTURE OVERVIEW

In Figure 1, a description of the proposed system to solve the Speaker diarization problem is presented. As it was already introduced, the first phase consists in extracting speech and non-speech segments from the audio raw signal. To do so, a Voice Activity Detection (VAD) module is needed. Subsequently, the signal is pre-processed in order to filter both background

or music noise as well as to compute some feature representations after performing an audio analysis. More specifically, the so-called Mel Spectrogram representation is calculated for this purpose. Then, these representations are passed through a supervised Deep Learning (DL) model based on Convolutional Neural Networks (CNN). This network is used to train a speaker classifier system which is capable of extracting latent representations of the speakers considering their Mel spectrograms as inputs. These latent representations are the ones denoted as embeddings. Finally, a Matching algorithm is used to match and associated an input embedding to the speaker who fits best according to a certain distance metric such as the cosine similarity distance.

Furthermore, the system implementation is slightly different when training or validating the model (development phase). More specifically, when training the model, the VAD module is not required since the data is already separated in speech segments and thus, this module is omitted during this phase. On the other hand, in a real situation, when a new audio signal is introduced into the system, the VAD module is employed to extract such relevant segments which correspond to speech information that needs to be both pre-processed and finally classified.

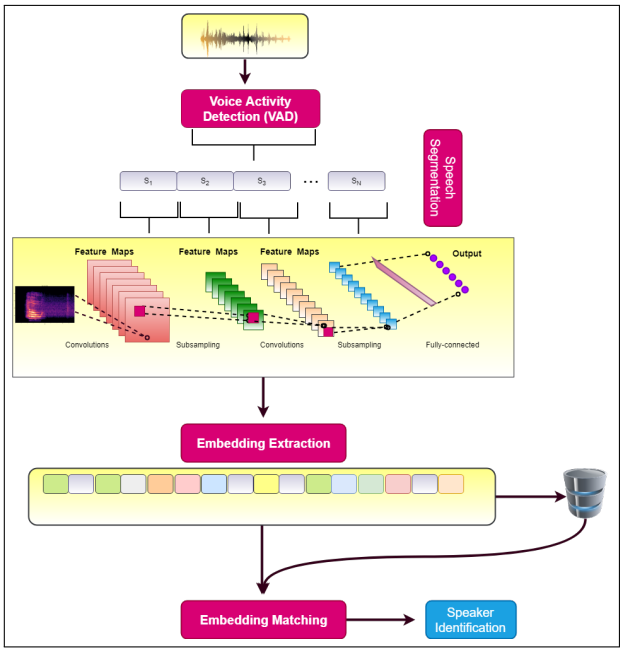

Fig. 1. A general block diagram of the proposed E2E architecture for Speaker Diarization.

A. Speaker Diarization Datasets

To perform the experiments, the LibriSpeech ASR corpus [5] was used. More specifically, the dataset contains a large volume of corpus of read English speech sampled at 16kHz. Therefore, the segmentation phase is not necessary to train the models due to the fact that each utterance contains only speech regarding one specific speaker. In 1, the results of the different experiments that were conducted are presented. As one may observe, the precision of the system in this specific dataset is very high since the data is well-prepared and there is no overlapped segments and or background noise.

On the other hand, we created different datasets using the real media content which was provided by some broadcasters partners from the Easy TV consortium. We focused on the one we denoted as CSFA after the acronym of a specific multimedia content provided by some of the partners of the consortium. The dataset was annotated in a semi-automatic fashion using a character annotation tool developed within the project. Using these annotations, the different audio segments associated with each character are extracted and stored together to make easier the extraction of the data during the training process. In this case, a preprocessing step is required to remove the background noise and music from the original audio signals. Moreover, this dataset is more realistic since it is based on real media content so that, some relevant problems may arise including overlapped sounds or speakers, or the problem of unbalanced data since some characters do not appear in many episodes.

B. Voice Activity Detector

The main goal of the VAD system consists in efficiently separate speech from non-speech audio segments. Several state-of-art models were both investigated and implemented to achieve the best performance in this first stage of the SDS pipeline. There exist different families of approaches: ones related to statistical modelling and others based on DL approaches. During the experiments conducted for this investigation, both families were analysed and developed to compare the performance in different scenarios. The first approach is based on changes in the energy within the human voice frequency band using median filters as authors describe in [6]. The second approach consists of a DL model that is trained to detect both gender and activity/non-activity events [7]. The architecture is based on CNN layers to extract relevant features from the audio signal and fully connected layers to perform the classification task. The third implemented approach is based on DL models as well but in this case, the output of the model is a binary decision which returns either speech or non-speech by using a pre-trained model developed within Google's Real-time communication for the web project.

Since the datasets used in this work already provided the required speech-segments, this stage is skipped when training the End-to-End (E2E) architecture.

C. Speaker Embedding Generation via CNN's

The third phase of the system consists in identifying the speaker of all the speech segments. To do so, the Mel spectrogram is computed for all the available samples in the dataset. More specifically, some basic parameters were defined including the windows size of the analysis which was fixed to 2048 samples, the Hop length was fixed to 512 and the number of filters was fixed as 96 after several testing experiments.

Then, a CNN model is trained to automatically identify the different speakers based on the Mel spectrogram input. The architecture of this model consists in a three-block-VGG-Based CNN with convolutional, max-pooling and fully-connected layers. Moreover, during the training process, the so-called categorical cross-entropy loss function is minimized using the Adam optimizer technique. Additionally, a cross-validation strategy is followed to prevent the model to overfit.

Once the model is trained, the last hidden layer is used as a compressed representation of the input which it is generally denoted as a speaker embedding. More specifically, such layer consists in a low-dimensional vector that better represents the Mel spectrogram introduced in the network. Finally, in order to generate all the speaker embeddings, the whole dataset is

Table 1. Comparison among the different experiments conducted by employing public and custom datasets for training the whole system.

Dataset name	*Domain*	*# Speakers*	*Accuracy (Train | | Val)*	*Precision (Train | | Val)*	*Recall (Train | | Val)*	*F1-Score (Train | | Val)*
LibriSpeech ASR corpus	public	7	96.38 | | 98.82	96.97 | | 99.06	95.35 | | 98.45	96.40 | | 98.76
LibriSpeech ASR corpus	public	20	86.84 | | 84.05	89.72 | | 87.68	84.58 | | 80.74	83.96 | | 84.06
CSFA	private	21	64.32 | | 63.68	65.25 | | 64.71	3.97 | | 64.03	64.60 | | 64.36

passed throughout the network and the set of vectors obtained in the embedding layer are stored to be used in the matching embedding stage.

D. Matching Speaker Embedding Algorithm

The main drawback when storing all the embeddings in the database lies in the fact that many of them may not be very representative and may lead to errors when predicting a new speaker. To address this problem, a clustering method is used to remove possible outliers: those embeddings associated to a certain speaker that may not represent it as expected. The proposed algorithm to address this issue is named as Density-Based Clustering Based on Hierarchical Density Estimates (HDBSCAN) [8] which is an improvement of the density-based, hierarchical clustering method proposed in [9]. By performing this approach for every set of embeddings associated to each speaker, we are removing from the storage those embeddings which are outliers of the speaker and normally are associated with overlapped segments, or false positive events. Additionally, the centroids of each of the classes (speakers) are also stored for being used when predicting a new audio signal based on the remaining embedding.

Furthermore, a distance-based algorithm is needed as well to perform the final classification task when a new audio signal is introduced into the system. More specifically, When a new input feeds the E2E, a new speaker embedding is obtained and thus, it must be associated to any of the speakers available in the dataset if and only if, the distance between such speaker embedding and the centroid embedding is lower than a particular threshold.

If all the distances between pairs of embedding centroids and the new speaker embedding do not satisfy the criteria, then, the new audio signal is automatically assigned to "New Speaker", otherwise it is associated to the speaker whose centroid is closer to the speaker embedding.

E. Experimental Results

In Table 1, the results of the experiments conducted during this investigation are presented using the classical metrics employed in classification tasks. As expected, the classification metrics are much better in the public datasets since the data is balance, there is no background noise or any other event that may overlap the voice of the speaker. On the other hand, when facing with more realistic datasets such as the one built during this investigation, the system suffers from a slight degradation which can be mitigated if the data is better annotated and by incorporating more preprocessing procedures to clean as much as possible the voice signal of the speakers.

3. CONCLUSIONS AND FUTURE WORK

This paper describes a complete deep learning architecture to automatically detect the speaker in any multimedia content throughout audio processing, CNNs and embedding matching algorithms.

The proposed solution is mainly focused on optimizing the subtitles generation process by automatically incorporate the current speaker in a certain moment of the multimedia which is a complex time-consuming tasks normally done by professionals of the multimedia industry.

Furthermore, the proposed schema takes advantage of CNNs in order to extract relevant features from voice signals which are generalized for any speaker. Finally, these features named as embeddings are then passed throughout a matching algorithm in order to both search and retrieve the target speaker.

Moreover, future research will be performed by merging the information from this pipeline together with some other image recognition system in order to better detect the current speaker as well as to add more value to the proposed architecture.

ACKNOWLEDGMENT

This work was supported by the H2020 European Project: Easy TV. Grant no. 761999.

REFERENCES

1. 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.
2. 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.
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. Z. Zajíc, M. Hrúz, and L. Müller, "Speaker diarization using convolutional neural network for statistics accumulation refinement." in *INTERSPEECH*, (2017), pp. 3562–3566.
5. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, 2015), pp. 5206–5210.
6. J. Pang, "Spectrum energy based voice activity detection," in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, (IEEE, 2017), pp. 1–5.
7. D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, 2018), pp. 5214–5218.

8. L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” J. Open Source Softw. **2**, 205 (2017).
9. R. J. Campello, D. Moulavi, and J. Sander, “Density-based clustering based on hierarchical density estimates,” in *Pacific-Asia conference on knowledge discovery and data mining,* (Springer, 2013), pp. 160–172.