Semi-supervised audio-driven TV-news speaker diarization using deep neural embeddings

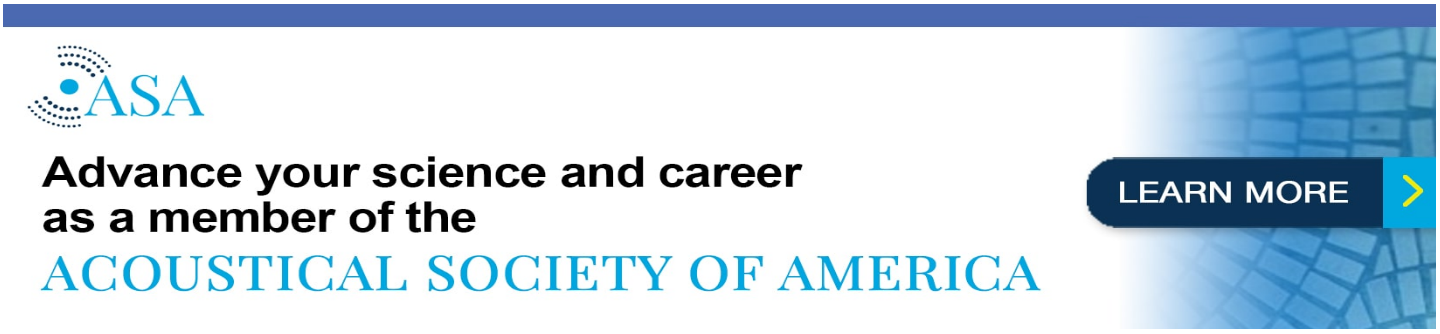
Nikolaos Tsipas, Lazaros Vrysis, Konstantinos Konstantoudakis, and Charalampos Dimoulas

Citation: The Journal of the Acoustical Society of America **148**, 3751 (2020); doi: 10.1121/10.0002924

View online: https://doi.org/10.1121/10.0002924

View Table of Contents: https://asa.scitation.org/toc/jas/148/6

Published by the Acoustical Society of America



JASA ARTICLE

Semi-supervised audio-driven TV-news speaker diarization using deep neural embeddings[a]

Nikolaos Tsipas,[b] Lazaros Vrysis,[c] Konstantinos Konstantoudakis,[d] and Charalampos Dimoulas[e]
Aristotle University of Thessaloniki, Thessaloniki, Greece

ABSTRACT:
In this paper, an audio-driven, multimodal approach for speaker diarization in multimedia content is introduced and evaluated. The proposed algorithm is based on semi-supervised clustering of audio-visual embeddings, generated using deep learning techniques. The two modes, audio and video, are separately addressed; a long short-term memory Siamese neural network is employed to produce embeddings from audio, whereas a pre-trained convolutional neural network is deployed to generate embeddings from two-dimensional blocks representing the faces of speakers detected in video frames. In both cases, the models are trained using cost functions that favor smaller spatial distances between samples from the same speaker and greater spatial distances between samples from different speakers. A fusion stage, based on hypotheses derived from the established practices in television content production, is deployed on top of the unimodal sub-components to improve speaker diarization performance. The proposed methodology is evaluated against VoxCeleb, a large-scale dataset with hundreds of available speakers and AVL-SD, a newly developed, publicly available dataset aiming at capturing the peculiarities of TV news content under different scenarios. In order to promote reproducible research and collaboration in the field, the implemented algorithm is provided as an open-source software package. © *2020 Acoustical Society of America.*

https://doi.org/10.1121/10.0002924

(Received 25 August 2020; revised 25 November 2020; accepted 25 November 2020; published online 15 December 2020)

[Editor: Bozena Kostek] Pages: 3751–3761

I. INTRODUCTION

During the last decade, an unprecedented, exponential growth of publicly available multimedia content has been observed, a trend that is expected to continue in the foreseeable future. Recent YouTube statistics reveal that over 300 h of video are uploaded to the streaming service every minute and over one billion hours of content are watched daily (YouTube). While recent developments in software tools and hardware infrastructure seem to provide the services and overall resources to support and sustain this level of growth (Aggarwal *et al.*, 2013; Zhu *et al.*, 2011), large-scale, automated semantic analysis of multimedia content remains a great challenge (Bello-Orgaz *et al.*, 2016). Nevertheless, research is ongoing on this emerging field, taking advantage of the associated progress in cloud computing and big data technologies (Yao *et al.*, 2015; Zhou *et al.*, 2016). From this perspective, algorithms and tools providing solutions around multimedia information retrieval (Benavent *et al.*, 2013; Nathwani *et al.*, 2013; Tsipas *et al.*, 2015b), multimodal speaker segmentation and tracking (Barnard *et al.*, 2014; Essid and Févotte, 2013; Minotto *et al.*, 2015; Nathwani *et al.*, 2013), content discovery (Fields *et al.*, 2011; Tsipas *et al.*, 2015a), recommendation (Zhao *et al.*, 2018), and categorization (Cho *et al.*, 2015) become critical contributors in the effort to keep the vast amounts of available multimedia assets accessible and relevant to the end user. In many of these cases, audio-driven processing offers computational and functional advantages, regarding multimodal event detection and segmentation, in complex audiovisual monitoring procedures (Dimoulas, 2016; Dimoulas and Symeonidis, 2015; Izadinia *et al.*, 2013; Stowell *et al.*, 2015).

In this context, speaker diarization, the process of partitioning an input audio stream into homogeneous segments based on the speaker identity, can provide the means to simplify content discovery and search procedures across large volumes of audiovisual information. Speaker diarization research has been applied to a number of areas over recent years, ranging from the processing of content from telephone communications (Shum *et al.*, 2013) and business meetings (Boakye *et al.*, 2008) to the analysis of radio and TV broadcasting streams (Barras *et al.*, 2006). Historically, most related approaches have been inspired by methodologies originating from the fields of speaker identification and verification. In accordance with this trend, Hidden Markov Model (HMM; Wooters and Huijbregts, 2008) techniques as well as Gaussian Mixture Models (GMM; Shum *et al.*, 2013) and I-vectors (Sell and Garcia-Romero, 2014) have been successfully utilized in speaker diarization applications based on the notion that a unique fingerprint, i.e., embedding, can be derived for every speaker. More recently, breakthroughs in the image processing field, demonstrated by the use of deep learning algorithms, have inspired the adoption of similar methods. In particular, Convolutional

[a] This paper is part of a special issue on Machine Learning in Acoustics.
[b] Electronic mail: nitsipas@auth.gr, ORCID: 0000-0001-7232-8839.
[c] ORCID: 0000-0003-2900-4657.
[d] ORCID: 0000-0001-5092-8796.
[e] ORCID: 0000-0001-7923-9361.

Neural Networks (CNN; Garcia-Romero *et al.*, 2017; Kumar *et al.*, 2020; Vrysis *et al.*, 2020; Vryzas *et al.*, 2020) and Recurrent Neural Networks (RNN; Wang *et al.*, 2017) have been employed to generate speaker embeddings usable in speaker diarization scenarios. In many cases, the above techniques have been developed in conjunction with similarity/distance–based methodologies (Bredin, 2016), initiating from the face recognition field (Schroff *et al.*, 2015).

Although a large subset of pioneering work in this area consists of unimodal methods relying on the analysis of the input audio signal, the potential gains of exploiting visual information, when available, have been demonstrated quite early in research (Otsuka *et al.*, 2008). Multimodal approaches can be coarsely classified into two main categories. The first exploits special hardware, including multiple cameras and microphones, with a predominant focus on the analysis of content produced in meetings (Gebru *et al.*, 2018). The second, emphasizing speaker diarization in TV talk-shows and news content, encompasses several techniques that do not require special hardware but rely on the semantic analysis of the available modes, mostly audio and image (Bost *et al.*, 2015; Bozonnet *et al.*, 2010; Friedland *et al.*, 2009; Noulas *et al.*, 2012; Vallet *et al.*, 2013). Thus, unimodal performance is improved by incorporating a multimodal fusion stage.

In this paper, a multimodal methodology for speaker diarization is proposed and evaluated in the context of TV news and talk shows. The audio analysis module is inspired by the work of Schroff *et al.* (2015) and Brendin (2016); however, the presented algorithm is optimized for reduced hardware requirements by employing a novel, Siamese long short-term memory (LSTM) network topology. The image processing module is optimized for re-use of existing pretrained models (He *et al.*, 2016; Sagonas *et al.*, 2016), based on the hypothesis that a naive exploitation of an additional mode of information can augment audio-based speaker diarization performance. The adopted approach can be applied on audiovisual content with single audio and visual tracks, without deploying any spatial sound localization. Based on this, the task can be considered more difficult compared to related multimodal approaches (Ban *et al.*, 2017; Gebru *et al.*, 2018).

The rest of the paper is organized as follows: in Sec. II, the proposed system is introduced and technical details of the audio, visual, and fusion sub-modules are provided. In Sec. III, the assessment procedure is outlined, and experimental results are presented and reviewed. Finally, in Sec. IV the overall contribution of this work is discussed and future research directions are proposed.

II. MULTIMODAL SPEAKER DIARIZATION

The proposed semi-supervised multimodal speaker diarization system consists of three processing stages (see Fig. 1). In stages A and B, two unimodal approaches are employed, one for each available mode (audio and visual), resulting in two segmentation instances of the analyzed audiovisual content. In stage C, a rule-based fusion of the preceding unimodal stages is performed to derive the final, multimodal speaker diarization output.

A. Audio-based processing (stage A)

The first processing module (stage A) focuses on the analysis of the audio content found in the input video. As illustrated in Fig. 1, the following steps are involved. A Voice Activity Detector (VAD) is employed to filter out audio content that does not contain speech. Subsequently, a feature-based representation of the audio signal is obtained, and the produced feature vector is fed into an LSTM neural network that generates embeddings in a multi-dimensional space. Finally, the generated mappings are clustered into M clusters, matching the number of speakers provided by the user,

$$S = \{1, 2, \ldots, M\}. \quad (1)$$

1. Voice activity detection

The aim of this step is to filter the original audio to minimize the amount of non-speech segments, thus producing a homogeneous/speech-only stream of data. The benefits are two-fold. First, the quality of the training data is improved as non-speech information is filtered out. Second, the elimination of non-speech data simplifies clustering configuration allowing to explicitly set the number of clusters equal to the number of speakers provided by the user. Since VAD is not the main focus of this work, and a well-controlled environment (TV news studio) is assumed for the deployment of the

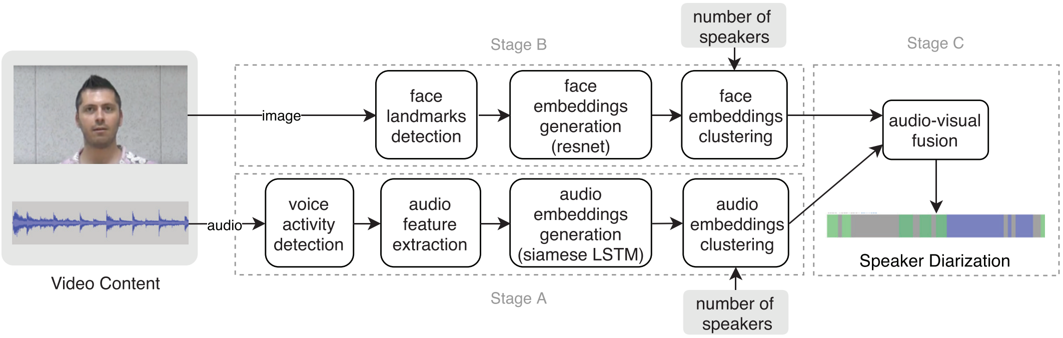


FIG. 1. (Color online) High level diagram of the proposed system where the three different processing stages are highlighted.

system, an off-the-shelf solution was chosen. The employed WebRTC VAD (Johnston and Burnett, 2012) is a Gaussian Mixture Model (GMM) based algorithm for real-time speech processing. The implementation is widely accepted as one of the modern and fast VAD options, and is available as an open-source software package.

2. Audio feature extraction

The audio feature extraction aims at generating data that can be utilized for the training of a binary classifier accepting sequences of vectors as input. Initially, as illustrated in Fig. 2, N_S feature vectors (V_F) are extracted from the input audio signal for every speaker. The feature vectors consist of $dim(V_F)$ components and are extracted using a window length of w_S samples and a step size of τ_S samples. Subsequently the extracted features vectors are aggregated to form successive sequences (Q^μ) of N_{agr} vectors. Due to the windowing process, overlapping sequences are produced, where the overlap is controlled by step size τ_L. The overall duration of a sequence is equal to $w_L = N_{agr} \cdot w_S$, whereas the total number of sequences per speaker is equal to N_L. A produced sequence of feature vectors is given in Eq. (2), where j denotes the index of a feature vector from speaker μ,

$$Q^\mu = \{V_F^\mu(i) | i \in \langle j, j+1, \dots, j+N_{agr}\rangle\},$$

where

$$j \in \langle 1, 2, \dots, N_S\rangle,$$
$$\mu \in S. \tag{2}$$

3. Siamese LSTM neural networks

A novel LSTM Siamese neural network approach is employed for the audio-based speaker diarization subsystem. The employment of LSTM networks is based on the hypothesis that their ability to selectively pass information across time can be beneficial in speaker recognition scenarios. Additionally, the use of a bidirectional network allows for the capture of more context as the output layer is able to get information from past (backward) and future (forward) states simultaneously. Whereas CNNs exhibit spatial structure learning properties (Vrysis *et al.*, 2020) and Time-Delay Neural Networks (TDNN) are shift-invariant (Myer and Tomar, 2018), LSTM networks are able to pass information across time and model long-term dependencies using gating mechanisms (Hochreiter and Schmidhuber, 1997; Zhang *et al.*, 2016). All three have been successfully employed in literature for temporal sequence classification, with no clear preference (Purwins *et al.*, 2019). From previous work of the research team in this domain (Vrysis *et al.*, 2020), it was concluded that CNNs exploit the spatial structure of input data and can be used for the classification of multivariate time-series data. Nonetheless, they are feed-forward networks without recurrent connections, hence the choice of LSTM allows the evaluation of this alternative, RNN-based approach. Following these remarks, indicative experiments were conducted with other architectures, whereas empirical trial and error observations validated the initial hypothesis regarding the adopted approach (however, the direct comparison of different network architectures is out of scope in this work).

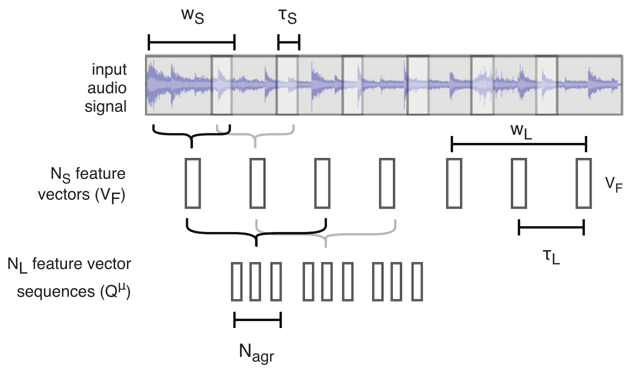

w_S
τ_S
input audio signal
w_L
N_S feature vectors (V_F)
V_F
τ_L
N_L feature vector sequences (Q^μ)
N_{agr}

FIG. 2. (Color online) Illustration of the hierarchical steps involved for the generation of the audio feature vector sequences (Q^μ) of a particular speaker (μ).

LSTM networks have demonstrated exceptional performance characteristics in sequence-to-sequence and sequence-to-vector classification scenarios, in a wide range of research areas (Chen and Wang, 2017; Illa and Ghosh, 2020; Sutskever *et al.*, 2014). For this particular problem, a sequence-to-vector approach is followed. The input *sequence* comprises a series of adjacent feature-vectors from a particular speaker, whereas the output *vector/embedding* (E_A) corresponds to a unique speaker identifier in a $dim(E_A)$-dimensional feature space.

Furthermore, a Siamese neural network topology is utilized to transform a multi-class classification problem into a binary one. As part of this methodology, two identical/sister LSTM networks with shared weights are employed, and the distance between their sequence-to-vector LSTM outputs (E_A) becomes the overall output of the network as illustrated in Fig. 3. As described in Eq. (3), during the training phase a pair (P_V) of feature vector sequences (Q_1^μ, Q_2^μ) from the same speaker (μ) or different speakers ($\mu i \neq \mu j$) is provided to the network, along with their relationship $Y \in [0, 1]$ (similar, dissimilar),

$$P_V = \begin{cases} \langle Q_1^\mu, Q_2^\mu\rangle, & \mu \in S, Y = 0, \\ \langle Q_1^{\mu i}, Q_2^{\mu j}\rangle, & \mu i, \mu j \in S, \mu i \neq \mu j, Y = 1. \end{cases} \tag{3}$$

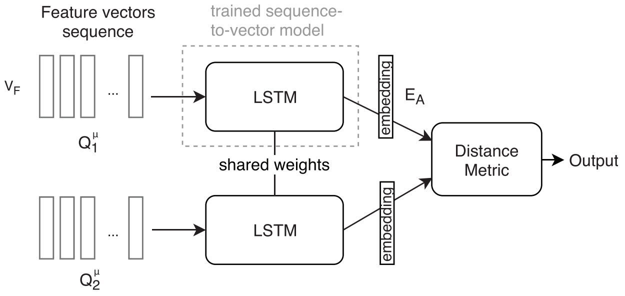


FIG. 3. The Siamese network topology employed for the training of the sister LSTM networks used for speaker embeddings generation.

By following a similarity-based approach for training, it is possible to generate a significantly higher number of training data from the original dataset and thus improve the accuracy of the model. Specifically, given a training dataset of M speakers with N_L samples per class (MN_L samples in total), it is possible to produce a dataset of $(M^2N_L^2 - MN_L)/2$ unique similar and dissimilar pairs.

The contrastive loss function (Hadsell *et al.*, 2006) is employed for the training of the Siamese neural network, whereas the Euclidean/L2 metric is used as the distance function, based on the successful application of the metric in similar research (Schroff *et al.*, 2015). The loss function is defined in Eq. (4), where D_L is the Euclidean distance between the outputs $G(Q_1^\mu)$ and $G(Q_2^\mu)$ of the sister neural networks as described in Eq. (5), Y is the binary label assigned to the pair of input vectors (0 for similar, 1 for dissimilar), and m_{th} is the margin. Dissimilar pairs contribute to the loss function only when their distance D_L is smaller than the threshold m_{th},

$$e_L = (1-Y)\frac{1}{2}D_L^2 + Y\frac{1}{2}\{\max(0, m_{th} - D_L)\}^2, \tag{4}$$

$$D_L = \sqrt{\{G(Q_1^\mu) - G(Q_2^\mu)\}^2}. \tag{5}$$

4. Audio-based embeddings clustering

The user-provided number of speakers (M) is exploited by the audio-based speaker diarization algorithm to map the generated embeddings into the corresponding clusters. The general purpose K-means algorithm, which has been successfully employed in diarization systems throughout the literature (Ben-Harush *et al.*, 2012; Dimitriadis and Fousek, 2017), was selected. This choice was also supported by the fact that the number of clusters is provided by the user and that the embeddings were generated through a training process employing a distance metric.

B. Image-based processing (stage B)

As illustrated in Fig. 1, the second processing module (stage B) focuses on the analysis of the visual information, having video frames as input. Face landmark detection is applied to extract a subsection of the image containing the face. The extracted part of the image is converted to a face embedding (E_V) using a pre-trained CNN, and, finally, a clustering algorithm is used to group the embeddings according to the number of speakers provided by the user.

1. Face embeddings generation

A Histogram of Oriented Gradients (HOG) estimation is performed against each frame to detect speaker faces. Since face detection is not the main focus of this work, HOG was primarily chosen because of the availability of an off-the-shelf algorithm provided as part of an open-source project (King, 2009). Furthermore, HOG has been successfully employed in face and human detection scenarios throughout the literature (Dalal and Triggs, 2005; Zhu *et al.*, 2006). Although its main disadvantage is sensitivity to image rotation (Cheon *et al.*, 2011), it is still a good candidate in the TV news domain where speakers are expected to face the camera most of the time.

Subsequently, a pre-trained 68-point landmark detection algorithm, developed by Sagonas *et al.* (2016), is used to extract landmarks from the frame containing the detected face. The landmarks, along with the frame, are provided as inputs to a pre-trained CNN able to generate corresponding embeddings for each input face. A pre-trained ResNet CNN with 27 layers (King, 2009) is used, essentially a variation of ResNet-34 developed by He *et al.* (2016). The generated embeddings represent a mapping of human faces to a $dim(E_V)$-dimensional space. Image frames of the same person are mapped near each other while image frames of different people are mapped further apart.

2. Face embeddings clustering

An approach similar to the one employed for the speaker embeddings categorization is followed. A semi-supervised clustering step is utilized to group the E_V embeddings according to the user-provided information regarding the number of speakers. Based on the ability of the system to identify the faces of the available speakers in the input video stream, each frame is annotated with the class (μ_V) of the detected face. The general purpose K-means algorithm is utilized in a way similar to the audio embeddings clustering. The resulting image-based segmentation is provided as input to the audio-visual fusion stage.

C. Audio-visual fusion (stage C)

The proposed audio-visual fusion attempts to exploit all the available modalities to enhance audio-driven speaker diarization performance. The employed methodology is based on the hypothesis that there is a one-to-one mapping between the appearance of a single face on a video frame and the actual speaker at that time interval. This observation, typical in TV talk-shows and news content, can be exploited to estimate a correlation between the outputs of the audio and visual processing modules in the time domain.

The first step in the fusion stage is the generation of a mapping between the faces and the speech samples produced by the different speakers. The image and audio-based segmentation outputs are partitioned using time frames with a length of w_F seconds, and an image/face class (μ_V) along with an audio/speech class (μ_A) is assigned on each frame. The total number of frames produced with window w_F is equal to N_{FA} for the audio modality and N_{FV} for the visual one. By iterating through all image frames (N_{FV}) containing the face of a particular speaker, the most frequently appearing speech class is calculated and mapped to that face. By the end of this procedure, a one-to-one mapping between M_V face classes and M_A speech classes ($M_V = M_A = M$) is derived, as described in Eq. (6)

$$f(\mu) = \underset{\mu_A}{\operatorname{argmax}} \left\langle \sum_{n \in N_{FV} | \mu_V = \mu} \delta(\mu_A^n, \mu_A) \right\rangle | \, \mu_A \in S. \tag{6}$$

Subsequently, a filtering step is employed to improve audio-based speaker diarization by exploiting visual information. When a mismatch between the face and the speech class is detected for a particular frame, then the fusion logic is triggered to calculate the fusion class (μ_F) of the frame. A rule-based, K nearest-neighbors approach is employed, where $K/2$ frames before and $K/2$ frames after the current position n are analyzed in order to detect the most frequently appearing face. This is achieved by creating a set of the appearing faces (I_n) in the K nearest frames, as described in Eq. (7), and afterward applying a conditional logic based on the cardinality ($|I_n|$),

$$I_n = \{\mu_V^k | k \in K_n\}$$

where

$$K_n = \left\langle n - \frac{K}{2}, \ldots, n-1, n+1, \ldots, n + \frac{K}{2} \right\rangle,$$
$$n \in \langle 1, 2, \ldots, N_{FV} \rangle. \tag{7}$$

The following cases are considered and captured in Eq. (8):

- If only one face appears in the K nearest frames ($|I_n| = 1$), then the audio class mapped to that face, is the produced fusion class μ_F^n.
- If more than one faces appear in the K nearest frames ($|I_n| > 1$), then the first most frequently appearing face ($\mu_{V_1}^n$) is selected as the fusion class μ_F, only when its frequency is greater by a factor of ρ from the second most frequently appearing one ($\mu_{V_2}^n$).
- Otherwise the fusion class μ_F^n of a particular window is set equal to the audio class μ_A^n,

$$\mu_F^n = \begin{cases} f(\mu_{V_1}^n), & \text{if } |I_n| \geq 1 \text{ and } g(\mu_{V_1}^n, \mu_{V_2}^n) > \rho, \\ \mu_A^n, & \text{otherwise}, \end{cases}$$

where

$$\mu_{V_1}^n = \underset{\mu_V}{\operatorname{argmax}} \left\langle \sum\nolimits_{k \in K_n} \delta(\mu_V^k, \mu_V) \right\rangle_{| \, \mu_V \in S},$$

$$\mu_{V_2}^n = \underset{\mu_V}{\operatorname{argmax}} \left\langle \sum\nolimits_{k \in K_n} \delta(\mu_V^k, \mu_V) \right\rangle_{| \, \mu_V \in S - \mu_{V_1}^n},$$

$$g(\mu_{V_1}^n, \mu_{V_2}^n) = \frac{\sum_{k \in K_n} \delta(\mu_V^k, \mu_{V_1}^n)}{\sum\limits_{k \in K_n} \delta(\mu_V^k, \mu_{V_2}^n)}. \tag{8}$$

The $K/2$ frames correspond to an aggregated time window t_k with a duration of $w_F \cdot K/2$ s. A simple scenario of a mismatched classification is presented in Fig. 4 where $K = 12$ and $\rho = 4$ are used for illustration purposes.

A rule-based approach was chosen for the late-fusion stage, as the main goal was to highlight the potential gains of integrating additional modalities in a naive way. This configuration allowed us to explain how decisions are made in a more straightforward way, in comparison to adaptive/machine-learning methodologies. Rule-based methodologies are not uncommon and still find application in a wide variety of problems (Korvel and Kostek, 2019).

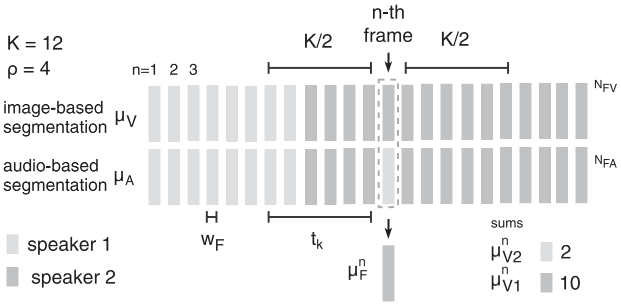


FIG. 4. Example scenario where a mismatch between the output of the image and audio processing modules is resolved as part of the fusion stage. Specifically, two speaker faces are detected in the $K = 12$ nearest frames around n, and since speaker 2 is the most frequently appearing face that satisfies $\rho = 4$, as described in Eq. (8), μ_F^n is set equal to $f(\mu_{V_1}^n)$.

III. EXPERIMENTAL RESULTS AND DISCUSSION

The goal of the conducted experiments is threefold. First, the accuracy of the audio-based binary classifier is evaluated along with the speaker diarization performance of the system using only the audio modality. Second, the image-based module is assessed regarding its ability to correctly identify the faces of involved parties. Finally, a set of experiments is conducted using the best-performing audio and (pre-trained) image models to assess the potential benefits of multimodal speaker diarization in comparison to unimodal (audio-only) speaker diarization. Furthermore, the results are evaluated against an existing state-of-the-art audio-based speaker diarization implementation.

The Voxceleb (Nagrani *et al.*, 2017) dataset is employed to train an audio-based model able to generate embedding vectors that can be used as unique speaker identifiers in a multidimensional space. The choice of this dataset was motivated by the high number of speakers provided, which can have a significant impact on a model's ability to generalize.

For the speaker diarization assessment, an ad-hoc audio-visual dataset developed for the requirements of this work is employed. Existing publicly available multimodal diarization datasets were evaluated; however, the most widely adopted ones (Czyzewski *et al.*, 2017; Gebru *et al.*, 2018) employ a multi-camera, multi-microphone setup, which does not align well with the scope of this work. The developed dataset is a simulation of a TV talk-show with 4 speakers ($M = 4$) in two versions with exactly the same audio content but different visual content. The first one (Test Video A) is a very "firm" video edit where the current speaker is always visible, whereas the second version (Test Video B) is more of a "loose" mix where the current speaker might not be always visible. The AVLab Speaker

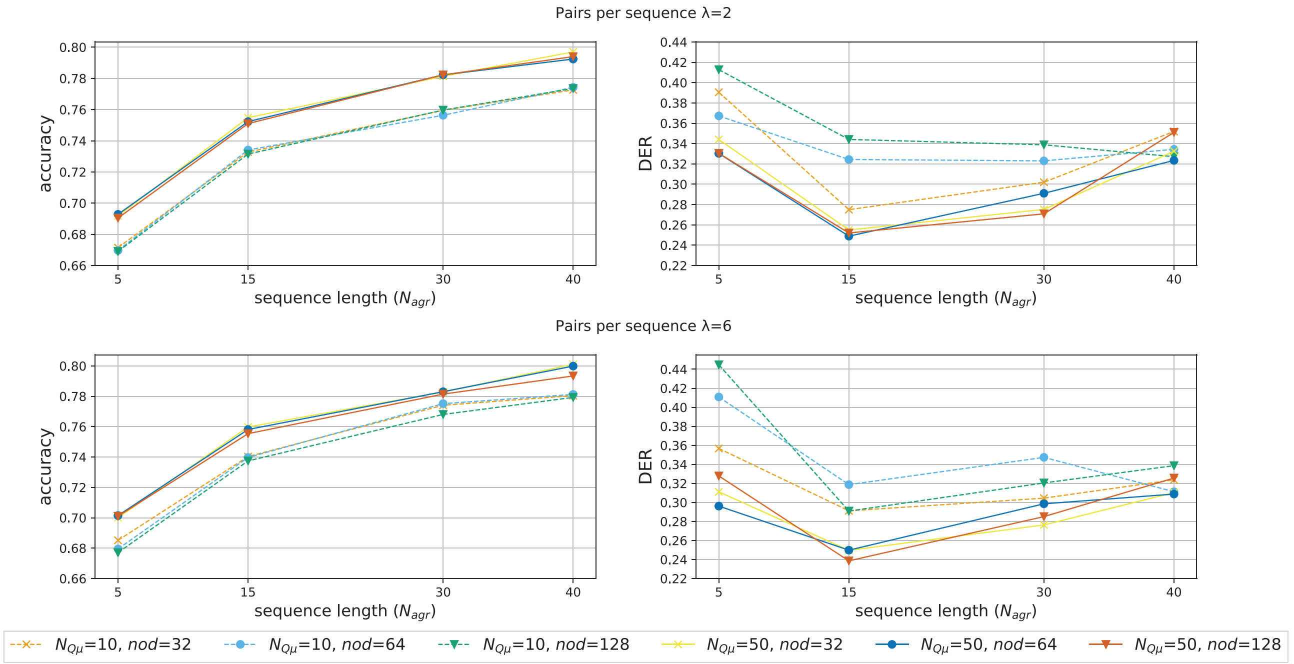


FIG. 5. (Color online) Siamese LSTM binary classification and audio-based speaker diarization results. The best performing model obtained using $n = 128$, $N_{Q\mu} = 50$, $\lambda = 6$, and $Nagr = 15$ is illustrated in the bottom right graph.

Diarization dataset[1] (AVL-SD) is made publicly available as part of this work in order to promote collaboration and reproducibility.

A. Audio modality configuration and evaluation

This subsection of the experiments consists mainly of a hyper-parameter optimization phase, where the configuration parameters for the audio-based binary classifier are selected using a grid search approach. For each set of evaluated parameters, the unimodal speaker diarization performance of the model is assessed at the same time. The K-means implementation available in the scikit-learn library (Pedregosa *et al.*, 2011) was selected, as more sophisticated variants did not provide any significant performance gains. As illustrated in Fig. 3, only one of the trained sister networks is needed in order to transform sequences of feature vectors into embedding vectors (E_A) during the speaker diarization evaluation. The accuracy metric is used for the binary classification performance evaluation, whereas the Diarization Error Rate (DER) is employed for the speaker diarization. The DER metric is calculated using the pyannote-metrics library (Bredin, 2017).

1. Audio dataset preparation

As already mentioned, the Voxceleb (Nagrani *et al.*, 2017) dataset is employed to train a model able to receive two sequences of feature vectors and produce a binary output, indicating whether the two sequences were generated from the same or different speakers. In total, speech audio from 1000 speakers is used for the development of the model, and the corresponding audio data is encoded as single channel, 16 bit, 16 000 Hz wav files. A number of audio files (ranging from 1 to 10) with recordings under different conditions comprise the available data for each one of the 1000 speakers/classes. No special handling of noise and reverberation is applied, as the proposed approach relies on the development of a model using a large and diverse dataset. Thus, as the model generalizes to adapt to the size and diversity characteristics of the training data (Gong *et al.*, 2019), the noise and reverberation characteristics of each recording become less relevant. A train/test split approach is followed where 90% of the speakers (900 speakers) is used for training and the remaining 10% (100 speakers) for testing. For each speaker, $N_{Q\mu}$ sequences out of the available N_L are randomly selected, and each sequence is paired: (a) with λ randomly selected sequences from the same speaker and (b) with λ randomly selected sequences from different speakers. The total number of input samples used to train the model is given in Eq. (9),

$$p_{train} = 900 \cdot N_{Q\mu} \cdot 2 \cdot \lambda. \tag{9}$$

The size of the test dataset is fixed throughout the experiments using $\lambda = 2$ and $N_{Q\mu} = 200$ as described in Eq. (10),

$$\begin{aligned} p_{test} &= 100 \cdot N_{Q\mu} \cdot 2 \cdot \lambda \\ &= 100 \cdot 200 \cdot 2 \cdot 2 \\ &= 80\,000. \end{aligned} \tag{10}$$

Since the performance evaluation of different audio features is not the main focus of this work, MFCCs with 20 components $[dim(V_F) = 20]$ are used, as similar configuration has been successfully employed in related research (Tsipas *et al.*, 2017). Likewise, typical windowing configuration employing a window length w_S of 1024 samples and no overlapping ($\tau_S = 1024$) was selected.

TABLE I. Siamese LSTM binary classification performance (audio-based).

			Accuracy		DER	
nod	N_{agr}	$N_{Q\mu}$	$\lambda = 2$	$\lambda = 6$	$\lambda = 2$	$\lambda = 6$
32	5	10	0.671	0.685	0.390	0.357
64	5	10	0.670	0.679	0.367	0.411
128	5	10	0.669	0.677	0.413	0.445
32	5	50	0.692	0.700	0.344	0.311
64	5	50	0.693	0.701	0.330	0.296
128	5	50	0.690	0.701	0.330	0.328
32	15	10	0.733	0.740	0.275	0.291
64	15	10	0.734	0.740	0.324	0.319
128	15	10	0.731	0.737	0.344	0.291
32	15	50	0.755	0.760	0.255	0.250
64	15	50	0.752	0.758	0.249	0.250
128	**15**	**50**	**0.751**	**0.755**	**0.252**	**0.239**
32	30	10	0.760	0.774	0.302	0.304
64	30	10	0.756	0.775	0.323	0.347
128	30	10	0.760	0.768	0.339	0.320
32	30	50	0.781	0.783	0.275	0.276
64	30	50	0.782	0.783	0.291	0.299
128	30	50	0.782	0.781	0.271	0.285
32	40	10	0.773	0.780	0.352	0.324
64	40	10	0.774	0.781	0.334	0.311
128	40	10	0.774	0.779	0.327	0.339
32	40	50	0.797	0.801	0.332	0.310
64	40	50	0.792	0.800	0.323	0.309
128	40	50	0.794	0.793	0.351	0.326

2. Hyper-parameter optimization

As illustrated in Table I, the evaluated parameters of the model are the length of the sequence of feature vectors (N_{agr}), the number of nodes used in the Bidirectional LSTM layers (*nod*), the number of sequences derived per speaker ($N_{Q\mu}$), and the number of similar and dissimilar pairs created per sequence, where two cases are assessed, $\lambda = 2$ and $\lambda = 6$. Greater values of λ were not assessed because of Graphics Processing Unit (GPU) memory constraints. A variable step size $\tau_L = 0.2 w_L$ is used, resulting in 80% overlapping between the generated feature vector sequences. By reviewing the results of the evaluation, illustrated in Fig. 6, it can be inferred that a higher number of derived sequences per speaker ($N_{Q\mu}$) has positive impact on the classification performance. The same positive impact is observed as the number of pairs per sequence (λ) increases. Furthermore, although there is a linear relationship between the classification performance and the sequence length (N_{agr}), the lowest DER value is achieved for $N_{agr} = 15$ ($w_L = 0.96$ s). This can be explained by the fact that a larger sequence length results in a coarser-grained segmentation and, subsequently, worse diarization performance. Finally, $nod = 128$ yields slightly improved results, especially when more training sequences per user are used, $N_{Q\mu} = 50$. Based on the above, the selected set of parameters is $nod = 128$, $N_{Q\mu} = 50$, $\lambda = 6$, and $N_{agr} = 15$. The architecture of the optimized sister neural network produced with the above parameters is presented in Fig. 7. Note that $nod = 128$ corresponds to 256 LSTM nodes as a bidirectional layer is used. Throughout these experiments a speaker embedding size [dim(E_V)] of 128 and a contrastive loss margin (m_{th}) equal to 1 were used as varying the values of these parameters had statistically insignificant impact on performance.

3. Model generalization/number of speakers

The impact of the number of speakers on the generalization ability of the model has also been evaluated. Using the set of parameters selected in Sec. III A 2 and illustrated in Fig. 5, three experiments employing 10, 100, and 900 speakers were conducted. In all three cases the test dataset remained the same as in the previous experiments. The model classification accuracy and the corresponding DER metrics are derived using the clustering approach described in Sec. II A 4 and presented in Table II. Both metrics improve as the number of speakers in the training dataset increases, i.e., accuracy monotonically increases while DER monotonically decreases.

The impact of the number of speakers used for training is also visually inspected in Fig. 6. The embeddings of the four speakers from the AVL-SD dataset are projected on a two-dimensional space using t-Distributed Stochastic Neighbor Embedding (TSNE; Maaten and Hinton, 2008).

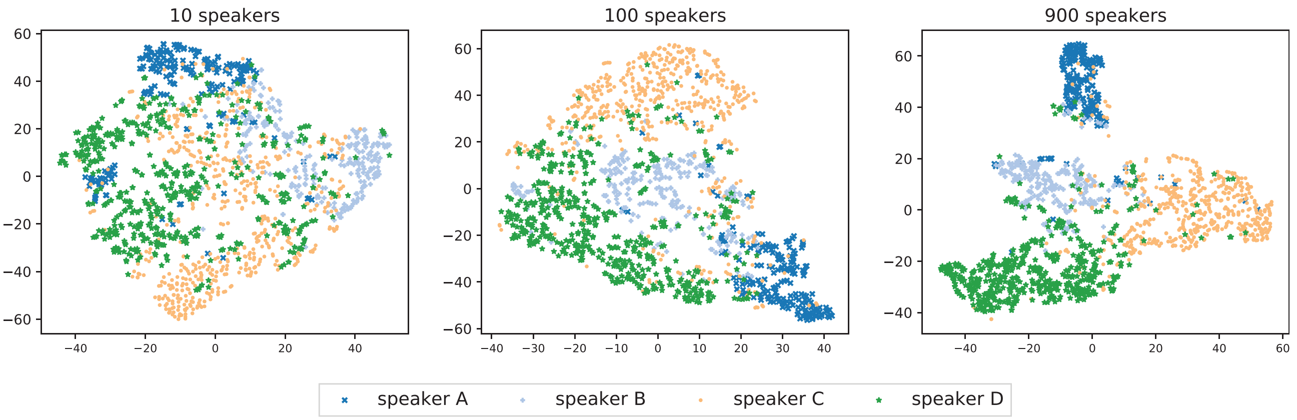


FIG. 6. (Color online) TSNE visualization to highlight the impact of the number of speakers available in the training dataset. As the number of speakers increases during the training phase, embeddings from the same speaker are projected closer to each other on the two-dimensional plane.

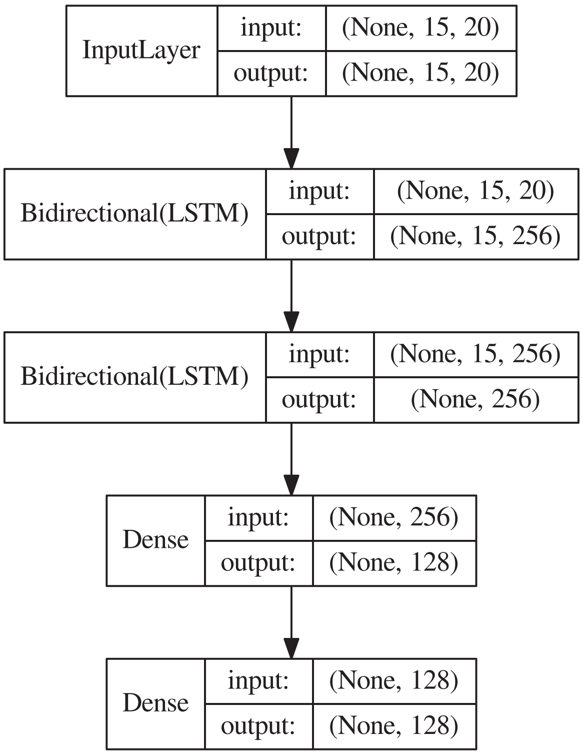


FIG. 7. Illustration of the different layers comprising the LSTM neural network with the best performing set of parameters. Specifically $nod = 128$, $N_{agr} = 15$, and $E_A = 128$ are used.

A perplexity value of 30 and 1000 iterations were chosen for the visualization through a trial-and-error approach. In all three subplots, the same number of vectors is projected using the above set of hyper-parameters. Although the cluster sizes and the distance between those clusters cannot be easily interpreted with TSNE, we observe that as the number of speakers increases embeddings from the same speaker are projected closer to each other on the two-dimensional plane.

B. Visual modality configuration and evaluation

The visual information is exploited by the image processing module in terms of detecting the face of a participating individual. Since the use of the visual modality on its own does not provide the ability to perform speaker diarization, the evaluation is focused on assessing its performance in accomplishing the task of identifying one or more faces presented on a video frame at any point in time. The different combinations of speakers appearing on the video frames are treated as different classes. In this dataset, the five classes appearing (as opposed to only four speakers) are *Speaker-1*, *Speaker-2*, *Speaker-3*, *Speaker-4*, and *All-speakers*, denoting frames where only one of the four speakers appears and frames where all four speakers appear. The ground truth data is generated from the two versions of the AVL-SD dataset using ELAN (Wittenburg *et al.*, 2006) and represents the reference segmentation. The hypothesis segmentation is the output of the visual module. The face-based segmentation error rate (FSER) is employed to measure the accuracy of the segmentation, which is calculated according to the standard DER metric, receiving as inputs the reference and hypothesis segmentations.

Input videos have a resolution of 640 × 360 pixels, and frames are extracted every 0.2 s. The selected sampling rate (5 frames per second) has been successfully employed in similar scenarios (Rowley *et al.*, 1998), and it is considered high enough to avoid temporal aliasing phenomena in TV talk-shows and news content. In order to perform landmark detection, a pre-trained 68-point landmark detection algorithm (Sagonas *et al.*, 2016) is employed, whereas face recognition is accomplished using a pre-trained ResNet CNN with 27 layers (King, 2009). The generated embeddings (E_V) consist of 128 components $[dim(E_V) = 128]$. Similarly to the audio processing module (stage B), the K-means algorithm is employed for the semi-supervised clustering of the generated embeddings. The evaluation results are presented in Table III, where it is clear that the image processing module is able to achieve a very low FSER for both versions of this particular dataset. Since one of the goals of this work was to assess the impact of additional modalities to the performance of the audio-driven speaker diarization algorithm, the high performance of the visual modality is considered a positive result. Thus, the evaluation of the multimodal improvements can be less biased, as a poor performing visual subsystem would have significant negative impact in overall performance.

C. Enhanced speaker diarization through multimodal fusion

The AVL-SD dataset, developed to simulate the peculiarities of TV news and talk-show content is employed in the overall assessment of the (end-to-end) multimodal system. In order to select an optimal set of parameters for the fusion stage, the algorithm is evaluated against both versions of the dataset. The parameters w_F, t_k, and ρ of the fusion stage are selected through a grid search approach, the results of which, are presented in Fig. 8. By reviewing the results, it becomes obvious that the best performance for both test videos is achieved for a frame resolution w_F of 50 ms, an aggregated window t_k of 2 s, and a factor ρ equal to 4. A smaller frame size ρ results in overtly better performance by enhancing the temporal accuracy of the multimodal diarization. Parameters ρ and t_k allow the fine tuning of the fusion

TABLE II. Impact of number of speakers on audio-based diarization.

No. speakers (training)	No. speakers (test)	Accuracy	DER
10	100	0.632	0.656
100	100	0.697	0.478
900	100	**0.755**	**0.238**

TABLE III. Face-based segmentation performance.

	Test Video A	Test Video B
FSER	0.047	0.093

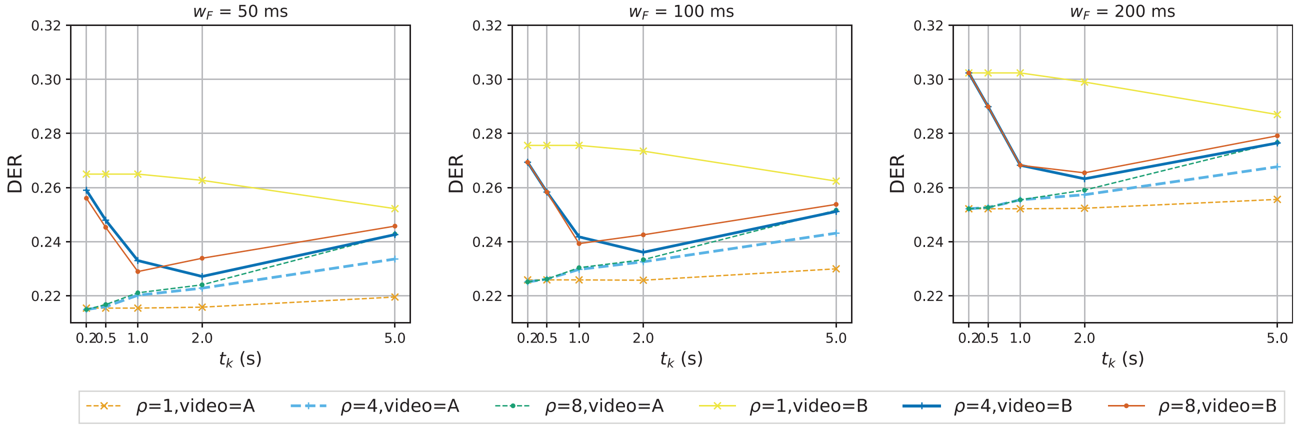


FIG. 8. (Color online) Fusion algorithm hyperparameter optimization for the frame resolution w_F, the aggregated window size t, and the factor ρ. The curves corresponding to a selected set of parameters $w_F = 50$, $t_k = 2$, and $\rho = 4$ are highlighted using thicker line width.

stage performance by controlling the contribution of the visual module to it.

The assessment results of the multimodal pipeline are presented in Table IV along with results from LIUM, an open-source, state-of-the-art, audio-based speaker diarization toolbox developed by Rouvier and Favre (2016) and the unimodal (audio-only) variant of the proposed algorithm.

By reviewing the evaluation results, the following observations can be made. The exploitation of visual information from multimedia content yields a DER improvement of 15% for the Test Video A of the AVL-SD dataset, whereas the corresponding improvement when the algorithm is evaluated against Test Video B is 11%. This is explained by the fact that Video A supports to a larger degree the main hypothesis on which the audiovisual fusion algorithm is based in comparison to Video B. Furthermore, the proposed algorithm yields a DER that is improved by a factor of two for the unimodal (audio-only) variant of the proposed algorithm and slightly over two for the multimodal (audio-visual) variant in comparison to the baseline LIUM algorithm.

IV. CONCLUSION AND FUTURE DIRECTIONS

In this paper, a multimodal speaker diarization algorithm employing deep neural network embeddings generated from audio and image content was introduced and evaluated. The advantages of a Siamese LSTM neural network topology for similarity-based training were illustrated as part of the speaker embeddings generation subsystem. The proposed unimodal speaker diarization system achieved superior performance in comparison to an existing state-of-the-art algorithm. Furthermore, the benefits of exploiting visual information to enhance audio-only speaker diarization performance were demonstrated by employing an approach optimized for reuse of existing pre-trained models. Finally, in order to promote reproducible research and collaboration in the field, the developed algorithm is made available as an open-source software package.[2]

TABLE IV. Unimodal/multimodal speaker diarization performance.

Algorithm	DER
baseline LIUM (audio-only)	0.513
proposed unimodal (audio-only)	0.258
proposed multimodal (Test Video A)	**0.222**
proposed multimodal (Test Video B)	0.231

Further research in this area could include the incorporation of the triplet loss function for the generation of the speaker and face embeddings, the employment of a training dataset with an even larger number of speakers and the assessment of hybrid, CNN-LSTM, or TDNN-LSTM approaches. Additionally, the introduced AVL-SD dataset could be extended to support diarization scenarios with a larger number of speakers, a feature that is missing from state-of-the-art audio-visual datasets. Finally, the image processing module could be enhanced by incorporating an RNN-based approach for speaker detection, focusing on the analysis of lip movement in the spatio-temporal space.

ACKNOWLEDGMENTS

The authors would like to thank Professor G. Papanikolaou for his supervision and guidance, and N. Vryzas, G. Boundourakis, and K. Liatsou for their contributions in the production of the AVL-SD dataset.

[1] https://github.com/avlab-auth/avlab-speaker-diarization-dataset

[2] https://github.com/avlab-auth/avlab-multimodal-speaker-diarization

Aggarwal, V., Gopalakrishnan, V., Jana, R., Ramakrishnan, K., and Vaishampayan, V. A. (**2013**). "Optimizing cloud resources for delivering IPTV services through virtualization," IEEE Trans. Multimedia **15**(4), 789–801.

Ban, Y., Girin, L., Alameda-Pineda, X., and Horaud, R. (**2017**). "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 446–454.

Barnard, M., Koniusz, P., Wang, W., Kittler, J., Naqvi, S. M., and Chambers, J. (**2014**). "Robust multi-speaker tracking via dictionary learning and identity modeling," IEEE Trans. Multimedia **16**(3), 864–880.

Barras, C., Zhu, X., Meignier, S., and Gauvain, J.-L. (**2006**). "Multistage speaker diarization of broadcast news," IEEE Trans. Audio Speech Lang. Process. **14**(5), 1505–1512.

Bello-Orgaz, G., Jung, J. J., and Camacho, D. (**2016**). "Social big data: Recent achievements and new challenges," Info. Fusion **28**, 45–59.

Benavent, X., Garcia-Serrano, A., Granados, R., Benavent, J., and de Ves, E. (**2013**). "Multimedia information retrieval based on late semantic fusion approaches: Experiments on a Wikipedia image collection," IEEE Trans. Multimedia **15**(8), 2009–2021.

Ben-Harush, O., Ben-Harush, O., Lapidot, I., and Guterman, H. (**2012**). "Initialization of iterative-based speaker diarization systems for telephone conversations," IEEE Trans. Audio Speech Lang. Process **20**(2), 414–425.

Boakye, K., Trueba-Hornero, B., Vinyals, O., and Friedland, G. (**2008**). "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4353–4356.

Bost, X., Linares, G., and Gueye, S. (**2015**). "Audiovisual speaker diarization of TV series," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4799–4803.

Bozonnet, S., Vallet, F., Evans, N., Essid, S., Richard, G., and Carrive, J. (**2010**). "A multimodal approach to initialisation for top-down speaker diarization of television shows," in *Proceedings of 18th European Signal Processing Conference*, pp. 581–585.

Bredin, H. (**2016**). "Tristounet: Triplet loss for speaker turn embedding," arXiv preprint arXiv:1609.04301.

Bredin, H. (**2017**). "Pyannote.metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *INTERSPEECH*, pp. 3587–3591.

Chen, J., and Wang, D. (**2017**). "Long short-term memory for speaker generalization in supervised speech separation," J. Acoust. Soc. Am. **141**(6), 4705–4714.

Cheon, M., Lee, W., Hyun, C.-H., and Park, M. (**2011**). "Rotation invariant histogram of oriented gradients," Intl. J. Fuzzy Logic Intel. Syst. **11**(4), 293–298.

Cho, K., Courville, A., and Bengio, Y. (**2015**). "Describing multimedia content using attention-based encoder-decoder networks," IEEE Trans. Multimedia **17**(11), 1875–1886.

Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., and Szykulski, M. (**2017**). "An audio-visual corpus for multimodal automatic speech recognition," J. Intel. Info. Syst. **49**(2), 167–192.

Dalal, N., and Triggs, B. (**2005**). "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1, pp. 886–893.

Dimitriadis, D., and Fousek, P. (**2017**). "Developing on-line speaker diarization system," in *Interspeech*, pp. 2739–2743.

Dimoulas, C. A. (**2016**). "Audiovisual spatial-audio analysis by means of sound localization and imaging: A multimedia healthcare framework in abdominal sound mapping," IEEE Trans. Multimedia **18**(10), 1969–1976.

Dimoulas, C. A., and Symeonidis, A. L. (**2015**). "Syncing shared multimedia through audiovisual bimodal segmentation," IEEE MultiMedia **22**(3), 26–42.

Essid, S., and Févotte, C. (**2013**). "Smooth nonnegative matrix factorization for unsupervised audiovisual document structuring," IEEE Trans. Multimedia **15**(2), 415–425.

Fields, B., Jacobson, K., Rhodes, C., d'Inverno, M., Sandler, M., and Casey, M. (**2011**). "Analysis and exploitation of musician social networks for recommendation and discovery," IEEE Trans. Multimedia **13**(4), 674–686.

Friedland, G., Hung, H., and Yeo, C. (**2009**). "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4069–4072.

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

Gebru, I. D., Ba, S., Li, X., and Horaud, R. (**2018**). "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," IEEE Trans. Pattern Analysis Mach. Intel. **40**(5), 1086–1099.

Gong, Z., Zhong, P., and Hu, W. (**2019**). "Diversity in machine learning," IEEE Access **7**, 64323–64350.

Hadsell, R., Chopra, S., and LeCun, Y. (**2006**). "Dimensionality reduction by learning an invariant mapping," in *Null*, IEEE, pp. 1735–1742.

He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Hochreiter, S., and Schmidhuber, J. (**1997**). "LSTM can solve hard long time lag problems," in *Proceedings of Advances in Neural Information Processing Systems*, pp. 473–479.

Illa, A., and Ghosh, P. K. (**2020**). "Closed-set speaker conditioned acoustic-to-articulatory inversion using bi-directional long short-term memory network," J. Acoust. Soc. Am. **147**(2), EL171–EL176.

Izadinia, H., Saleemi, I., and Shah, M. (**2013**). "Multimodal analysis for identification and segmentation of moving-sounding objects," IEEE Trans. Multimedia **15**(2), 378–390.

Johnston, A. B., and Burnett, D. C. (**2012**). *WebRTC: APIs and RTCWEB Protocols of the HTML5 Real-Time Web* (Digital Codex LLC).

King, D. E. (**2009**). "Dlib-ml: A machine learning toolkit," J. Mach. Learn. Res. **10**(Jul), 1755–1758.

Korvel, G., and Kostek, B. (**2019**). "Discovering rule-based learning systems for the purpose of music analysis," in *Proceedings of Meetings on Acoustics 178ASA*, Acoustical Society of America, Vol. 39, p. 035004.

Kumar, M., Kim, S. H., Lord, C., and Narayanan, S. (**2020**). "Improving speaker diarization for naturalistic child-adult conversational interactions using contextual information," J. Acoust. Soc. Am. **147**(2), EL196–EL200.

Maaten, L. v. d., and Hinton, G. (**2008**). "Visualizing data using t-SNE," J. Mach. Learn. Res. **9**(Nov), 2579–2605.

Minotto, V. P., Jung, C. R., and Lee, B. (**2015**). "Multimodal multi-channel on-line speaker diarization using sensor fusion through SVM," IEEE Trans. Multimedia **17**(10), 1694–1705.

Myer, S., and Tomar, V. S. (**2018**). "Efficient keyword spotting using time delay neural networks," arXiv preprint arXiv:1807.04353.

Nagrani, A., Chung, J. S., and Zisserman, A. (**2017**). "VoxCeleb: A large-scale speaker identification dataset," arXiv preprint arXiv:1706.08612.

Nathwani, K., Pandit, P., and Hegde, R. M. (**2013**). "Group delay-based methods for speaker segregation and its application in multimedia information retrieval," IEEE Trans. Multimedia **15**(6), 1326–1339.

Noulas, A., Englebienne, G., and Krose, B. J. (**2012**). "Multimodal speaker diarization," IEEE Trans. Pattern Analysis and Mach. Intel. **34**(1), 79–93.

Otsuka, K., Araki, S., Ishizuka, K., Fujimoto, M., Heinrich, M., and Yamato, J. (**2008**). "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization," in *Proceedings of the 10th International Conference on Multimodal Interfaces*, ACM, pp. 257–264.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.* (**2011**). "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**(Oct), 2825–2830.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (**2019**). "Deep learning for audio signal processing," IEEE J. Selected Topics Signal Process. **13**(2), 206–219.

Rouvier, M., and Favre, B. (**2016**). "Investigation of speaker embeddings for cross-show speaker diarization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5585–5589.

Rowley, H. A., Baluja, S., and Kanade, T. (**1998**). "Neural network-based face detection," IEEE Trans. Pattern Analysis Mach. Intel. **20**(1), 23–38.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., and Pantic, M. (**2016**). "300 faces in-the-wild challenge: Database and results," Image Vision Comput. **47**, 3–18.

Schroff, F., Kalenichenko, D., and Philbin, J. (**2015**). "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823.

Sell, G., and Garcia-Romero, D. (**2014**). "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *Spoken Language Technology Workshop*, pp. 413–417.

Shum, S. H., Dehak, N., Dehak, R., and Glass, J. R. (**2013**). "Unsupervised methods for speaker diarization: An integrated and iterative approach," IEEE Trans. Audio, Speech, Lang. Process. **21**(10), 2015–2028.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., and Plumbley, M. D. (**2015**). "Detection and classification of acoustic scenes and events," IEEE Trans. Multimedia **17**(10), 1733–1746.

Sutskever, I., Vinyals, O., and Le, Q. V. (**2014**). "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, pp. 3104–3112.

Tsipas, N., Vrysis, L., Dimoulas, C., and Papanikolaou, G. (**2017**). "Efficient audio-driven multimedia indexing through similarity-based speech/music discrimination," Multimedia Tools Appl. **76**(24), 25603–25621.

Tsipas, N., Vrysis, L., Dimoulas, C. A., and Papanikolaou, G. (**2015a**). "Content-based music structure analysis using vector quantization," in *Proceedings of the Audio Engineering Society Convention* 138.

Tsipas, N., Zapartas, P., Vrysis, L., and Dimoulas, C. (**2015b**). "Augmenting social multimedia semantic interaction through audio-enhanced web-tv services," in *Proceedings of the Audio Mostly on Interaction with Sound*, pp. 1–7.

Vallet, F., Essid, S., and Carrive, J. (**2013**). "A multimodal approach to speaker diarization on TV talk-shows," IEEE Trans. Multimedia **15**(3), 509–520.

Vrysis, L., Tsipas, N., Thoidis, I., and Dimoulas, C. (**2020**). "1d/2d deep CNNs vs. temporal feature integration for general audio classification," J. Audio Engineering Society **68**(1/2), 66–77.

Vryzas, N., Tsipas, N., and Dimoulas, C. (**2020**). "Web radio automation for audio stream management in the era of big data," Information **11**(4), 205.

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., and Moreno, I. L. (**2017**). "Speaker diarization with LSTM," arXiv preprint arXiv:1710.10468.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (**2006**). "ELAN: A professional framework for multimodality research," in *5th International Conference on Language Resources and Evaluation*, pp. 1556–1559.

Wooters, C., and Huijbregts, M. (**2008**). "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans* (Springer), pp. 509–519.

Yao, S., Wang, Y., and Niu, B. (**2015**). "An efficient cascaded filtering retrieval method for big audio data," IEEE Trans. Multimedia **17**(9), 1450–1459.

YouTube "Press Statistics," https://www.youtube.com/yt/about/press/, accessed 2018-08-16.

Zhang, H., Bao, F., Gao, G., and Zhang, H. (**2016**). "Comparison on neural network based acoustic model in Mongolian speech recognition," in *International Conference on Asian Language Processing*, IEEE, pp. 1–5.

Zhao, Z., Yang, Q., Lu, H., Weninger, T., Cai, D., He, X., and Zhuang, Y. (**2018**). "Social-aware movie recommendation via multimodal network learning," IEEE Trans. Multimedia **20**(2), 430–440.

Zhou, P., Zhou, Y., Wu, D., and Jin, H. (**2016**). "Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks," IEEE Trans. Multimedia **18**(6), 1217–1229.

Zhu, Q., Yeh, M.-C., Cheng, K.-T., and Avidan, S. (**2006**). "Fast human detection using a cascade of histograms of oriented gradients," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* Vol. **2**, pp. 1491–1498.

Zhu, W., Luo, C., Wang, J., and Li, S. (**2011**). "Multimedia cloud computing," IEEE Signal Process. Mag. **28**(3), 59–69.