|  |
| --- |
| The University of Hong Kong  Faculty of Engineering  Department of Computer Science  COMP7704  Dissertation Title  Real-time Speaker Recognizer  Submitted in partial fulfillment of the requirements for the admission to the degree of Master of Science in Computer Science  By  Pan Hao  3035349015  Supervisor's title and name: Dr. Beta C.L. Yip  Date of submission: 01/07/2019 |

**Abstract**

Speaker diarization has become more important in many speech processing tasks. Most state-of-the-art speaker diarization system decodes in an offline fashion and requires intensive computation and long processing time, which leads to the handicap for real-time applications. In this paper, we implemented the binary key speaker modelling approaches and built a fast offline speaker diarization system that can label the speakers in recorded audio, with visualization and audio player panel. An advanced version of the system is also developed, which can process in real-time with acceptable delay and correct earlier outputs when necessary.

Keywords: Speaker diarization, binary key speaker modelling, speaker clustering, MFCC

***The abstract should not be more than 200 words.***

**Declaration**

Table of Contents

[Introduction 3](#_Toc12633070)

[Analysis of problem 7](#_Toc12633071)

[Theoretical principles 8](#_Toc12633072)

[Method of investigation 8](#_Toc12633073)

[Design and construction of software system 8](#_Toc12633074)

[Design of the speaker diarization system 8](#_Toc12633075)

[Design of the Visualization Panel 14](#_Toc12633076)

[Theoretical/Algorithmic/Experimental results 16](#_Toc12633077)

[Discussion/Analysis of approach/results 16](#_Toc12633078)

[Conclusions 16](#_Toc12633079)

[Bibliography 20](#_Toc12633080)

# Introduction

1. ***the subject matter and the scope of the investigation,***
2. ***the purpose of the dissertation,***
3. ***the organisation of the report.***

***If applicable, a brief survey of previously published work and current trends may be included in this section.***

There is a growing need of using audio processing technologies to index, search, access and analyze the information from audio streams. In many real cases in the presence of multiple speakers including conversations, meetings, conference, broadcast news and debates, there are multiple audio sources or multiple speakers speaking within one audio channel. Speaker diarization, as the focus of audio diarization, is the process used in these cases to determine the number of speakers and assign speech segments to the corresponding speakers. Therefore, it is often summarized as “who spoke when” question [1]. Speaker diarization is an important front-end tool and the audio information output can be more efficiently used as input in searching and indexing audio archives, automatic speaker recognition and natural language processing.

The traditional three primary application domains of speaker diarization is broadcast news, recorded meetings and telephone conversation [2]. The audio streams from these domains are different in style of the speech, style of the noise source, numbers and locations of microphones, configuration of environment and therefore present unique challenges. [3] makes detailed comparison between broadcast news and conference meetings. The majority of the literature speaker diarization will only focus on one use cases and some propose specific techniques to tackle the unique challenge. For instance, [4] propose the acoustic beamforming technology to take advantage of the multiple microphones available in a meeting room domain to facilitate the speaker diarization process.

Speaker diarization can also be referred as speaker segmentation and clustering, as the majority of diarization approach consist two main steps of segmentation and clustering [5]. Speaker segmentation or speaker change detection (SCD) in many literatures, aims at finding the speaker change points in the audio so that splitting the original audio stream into speaker homogeneous segments [2, 6] . One popular segmentation algorithms is the use of Bayesian information criterion (BIC) firstly introduced by Schwarz [14] and firstly used in speaker segmentation by Chen and Gopalakrishnan [15]. Many state-of-art systems incorporate BIC as a segmentation metric in the following [16, 17]. As BIC approach is computationally intensive, several works (e.g. [18]) propose modification or other technologies used with BIC to speed up the process. Some common alternative segmentation approaches include Generalized Likelihood Ratio (GLR) [19] and Kullback–Leibler (KL) divergence [20]. Some recent papers propose advanced technology like deep neural network (DNN) to find speaker change points [21].

Clustering, on the other hand, focus on agglomeration of segments from segmentation step into groups that from the same speaker. One of the popular approaches recently is unsupervised i-vector clustering. Sell & Garcia-Romero [22] propose a system that uses i-vectors and probabilistic linear discriminant analysis (PLDA) which has good performance for multi-language telephone conversation data.

Speaker diarization system can be differentiated as offline and online system. The offline system have access the whole audio recording before processing, and the clustering step is performed only when complete audio stream has been segmented. This means the offline diarization system cannot be used in real-time applications where the analysis on the audio has to be conducted simultaneously or with acceptable latency when the audio is created. Online diarization, on the other hand, only have access to the audio data up to the point that is been recorded, which means the diarization have to perform in a “left-to-right” fashion [7] that process and assign the segments once they are created and detected in the audio stream. Therefore, online speaker diarization or real-time speaker diarization can be used in real-time applications like multi-person / human-computer voice interactive systems. Offline speaker diarization is the main focus in the field of speaker diarization [3] and there is limited work on online speaker diarization. [8] presents a real-time speaker diarization system for the meeting environment. However, the system relies on the speaker seat locations and has the limitation of detecting only one speaker in one frame even if there are multiple speakers speaking. [9] propose an online speaker diarization based on Gaussian mixture models (GMMs) and start the system with male, female and noise models, tested with broadcast news data. However, their system has difficulties dealing with speech overlapped by music. [7] propose a novel Maximum a Posteriori (MAP) adapted transform within the i-vector speaker diarization framework, which have good diarization result for English-speaking telephone conversation data.

# Analysis of problem

# Theoretical principles

# Method of investigation

# Design and construction of software system

## Design of the speaker diarization system

**Voice Activity Detection**

Voice activity detection (VAD) or speech activity detection (SAD) is included in the majority of the speaker diarization system. It is used to identify the region of the audio data that is being voiced or containing speech, from the unvoiced or non-speech regions that contain silence or background noise. Energy-based voice detection and model-based voice detection are two main approaches of VAD. The energy-based voice detection removes the silence part based on the energy level and has the advantage of simplicity and speed. However, this approach fails to distinguish the load noise from the speech, and therefore is ineffective in many application domains of speaker diarization [10, 11]. To avoid the limitation of energy-based detection, model-based detection that is developed on the different acoustic phenomena, are more frequently used in speaker diarization system [12].

The model-based VAD algorithm that Google developed for the WebRTC project [13] will be applied in the design of our system, the reason includes:

1. The WebRTC project targets on real-time capabilities and its VAD algorithm has been widely used for different delay-sensitive scenarios [14], which is suitable for our offline and real-time applications.
2. The free and open-source implementation of the algorithm is available and there is Python interface (Py-webrtcvad [15]) that is well-suited for our development.

The VAD process in the system can be performed on the all audio data for the audio recording or the new input data for real-time audio stream. The result of the VAD is a mask vector that used to exclude the data segments or the acoustic features that contains no speech from the original data.

**Feature extraction**

The raw audio data is usually converted into a sequence of acoustic feature vectors that contains speaker specific information before the segmentation and clustering. This feature extraction step try to get the acoustic features that contain formant information, model the mode of excitation and/or the shape of the vocal tract when people producing speech [12]. Common features include Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Codes (LPC), Linear Prediction Cepstral Coefficients (LPCC) [16] and constant Q transform Mel-frequency Cepstral Coefficients (ICMC). These features are different in frequency analysis and frequency smoothing techniques.

In this dissertation project, the focus will be put on the MFCCs which are frequently used in the community of speaker diarization and more specifically, the online speaker diarization system [9, 7]. Librosa, the python library for music and audio analysis [17], will be used to extract the MFCCs. The pseudo code is as below:



**Segmentation**

Speaker segmentation or speaker change detection (SCD) in many literatures, aims at finding the speaker change points in the audio so that splitting the original audio stream into segments that contains only one speaker [2, 6]. Although many system will used various distance metric to decide whether the speech of the adjacent windows are from the same or different speakers, a simple method that use fixed-sized windows to splits the audio into small equal-sized segments, will be used in our system.

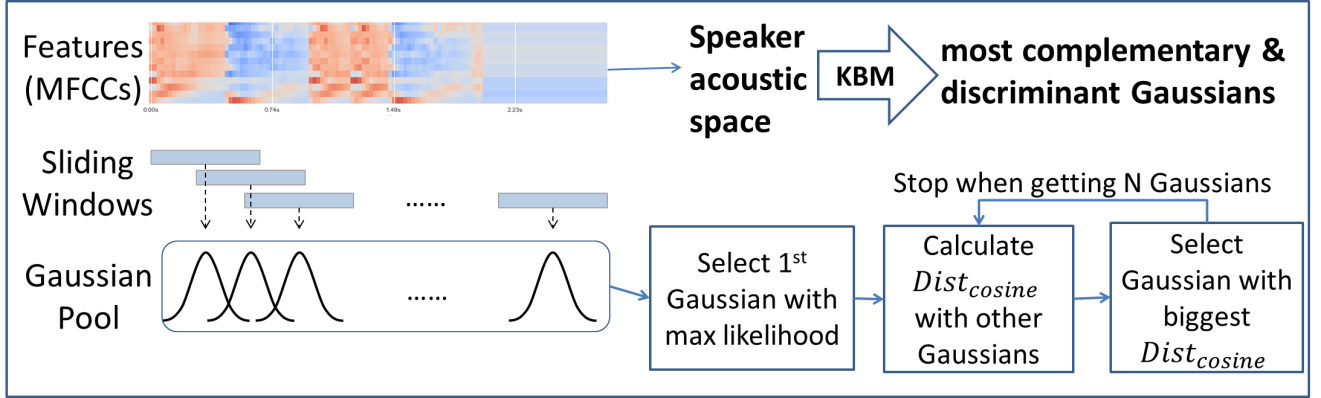
**Speaker Modelling**

The speaker modelling is the heart of the many state-of-art speaker diarization system, and the similar step is called embedding extraction in some other research. This step aims to find speaker-discriminative identifiers that can be used to distinguish unique speaker from each other. Some models or embeddings are first found to be effective in speaker recognition or identification tasks, and then are adopted in speaker diarization tasks. Famous speaker models used in state-of-art system include Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) and famous speaker embeddings include speaker factors [18], i-vectors [7] and d-vectors [19].

In this project, we will adopt the Binary Key (BK) speaker modelling techniques that is firstly proposed in [20] and later used in speaker diarization context in [21]. This technique has advantages of being domain-robust, requiring no external data and running faster than real-time, so therefore can be utilized for our offline and real-time system.

**Binary Key Background Model (KBM) Training**

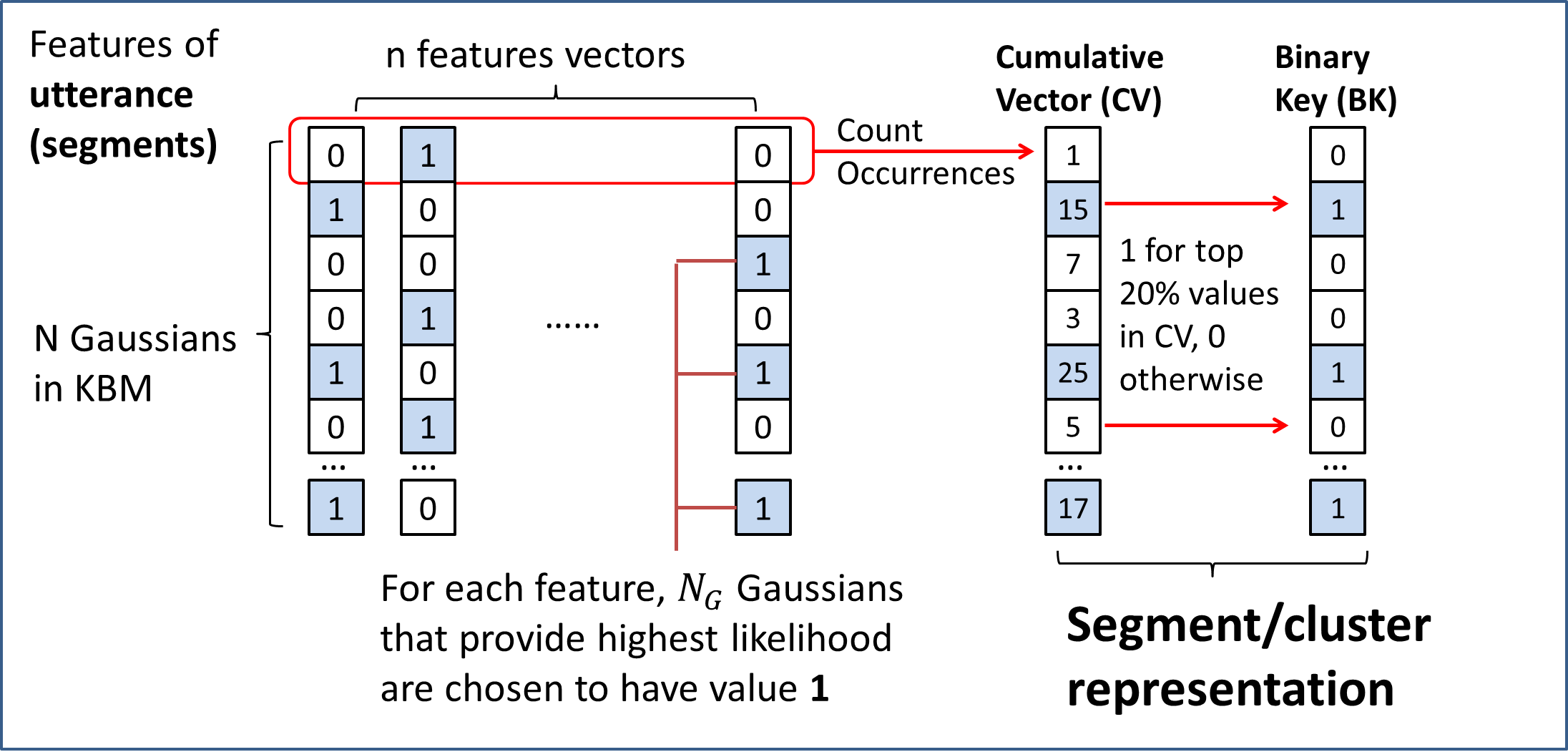
The first step of the binary key speaker modelling is to train a GMM-like model called KBM from the acoustic features (MFCCs for our system) extracted from the input data. While the majority of the speaker modelling or speaker embedding extraction techniques requires large amount of external training data that is from the same domain of the tested data, the BK techniques train the KBM directly from the input tested data. Therefore, BK modelling is domain-robust and can avoid the negative impacts from the mismatch of the acoustic conditions between training and tested datasets.

*Figure 1* shows the process of KBM training. To obtain the KBM, multiple Gaussians are trained on the data that is separated by a 2-second sliding window on the input features data. The mean and the standard deviation of the 2-second data will be calculated to get the Gaussian, which will be realized by multivariate\_normal function from scipy library in the script. The shifting rate of the sliding window is determined automatically to have enough number of Gaussians. All resulted Gaussians form a Gaussian pool that covers all acoustic space in the input audio. Then *N* Gaussians are selected from the pool using single-linkage clustering strategy with cosine similarity as similarity measures. The first Gaussian is selected with maximum likelihood and the Gaussians with highest dissimilarity with previously selected Gaussians is selected subsequent until getting *N* Gaussians. Then the KBM is formed by these *N* Gaussians, which are considered the most complementary and discriminat Gaussians that can be used to represent the speaker acoustic space. The number *N* can be a fixed number or be determined relatively by a percentage of the total number of Gaussians in the pool. 

*Figure 1 KBM Training*

**Feature binarization**

The second step of the binary key speaker modelling is the feature binarization that transforms feature vectors of an utterance to into a binary key, and this process is shown in Figure 2.

****

*Figure 2 Feature Binarization*

Firstly, a matrix with number of rows equal to *N* (size of KBM) and number of columns equal to number of feature vectors are created. For each column of each feature vector, rows representing the Gaussians that provide highest likelihood for the given features are chosen to have value 1. All other rows in this column are set to have value 0. This process recorded the best Gaussians from the KBM that represent each features. Secondly, we sum the values of each row, to form a vector called Cumulative Vector (CV) of size *N*. This CV contains the count of the occurrences of each Gaussians has been selected as the top-likelihood Gaussian for all feature vectors, so therefore larger number in the CV indicates higher impact of the corresponding Gaussian components

## Design of the Visualization Panel

A visualization panel for the speaker diarization / recognizer system can help the users to understand and examine the diarization results intuitively. The target of the design of the visualization panel includes:

1. To show the number of the speakers
2. To distinguish different speakers and their speech on the timeline
3. To allow the users to playback the audio to compare with the diarization results
4. To allow the users to choose the start point at the timeline to play the audio and pause it in anytime
5. To use the same programming language as the speaker diarization system

To complete the above targets of the visualization panel, two python modules viewer.py and player.py are designed respectively.

The player.py is the module to open, play and pause the input audio files.

The viewer.py will be used to show diarization results of the input audio or audio stream. The x-axis is the timeline of the audio while the y-axis shows the number of the speakers. Rectangle of different colors will be used to display different speakers and their speech in the timeline. The position and the length of the rectangles will be determined by the position and the length of the speech in the timeline respectively. The Matplotlib library, which is the most popular library in Python for 2D plotting will be used in this part.

# Theoretical/Algorithmic/Experimental results

# Discussion/Analysis of approach/results

# Conclusions

In the Discussion and Conclusions sections, critical evaluation of the techniques employed and results obtained should be carried out. Observations derived from the results should be compared with theoretical predictions. The conclusions should follow logically from the argument and results presented in the report. Recommendations for further investigations may also be included. Supplementary information not essential to the report's main thesis is best included under the heading of Appendices

1. **Objectives**

As the limited work on online diarization system and the importance of the speaker diarization for real-time applications, the objective of this dissertation is to build a real-time speaker system that fulfills the following requirements:

1. The system can perform speaker diarization for recorded audios like radio talk or phone conversation;
2. The system can perform speaker diarization in a real-time fashion, that can process the live speech audio and generate output as the input is analyzed;
3. The system should be language-independent and operating-system independent;
4. The system should not require the number of the speakers, identity of the speakers or the voice samples of the speakers for the training.
5. **Methodology**
   1. **Data**

Several databases that the audio recordings are transcribed into speaker segments, are available for training and testing in the development of speaker diarization system.

For meeting environment, sample databases include:

* ICSI Meetings Recorder corpus [22]
* NIST Meeting Pilot Corpus Speech [23]

For telephone conversation, sample databases include:

* LDC CALLHOME English corpus [24] and CALLHOME corpus in other languages are also available.

For broadcast news, sample databases include:

* LDC 1996 radio broadcast news database (HUB4) [25] and an overview of broadcast news corpora is conducted by Graff [26].
  1. **System design**

1. Data Preprocessing:

The first step of a prototypical speaker diarization system is the audio data preprocessing, which usually include noise-reduction, parameterization of speech data into acoustic features and speech activity detection (SAD) [6]. The audio data processed in this step will be the input of the following segmentation and clustering

The common features extracted for speaker diarization include:

* Mel Frequency Cepstral Coefficients (MFCC)
* Linear frequency cepstral coefficients (LFCC)
* Perceptual Linear Predictive (PLP)
* Linear Predictive Coding (LPC) [3].

In this dissertation project, the focus will be put on the MFCCs which are frequently used in online speaker diarization system [9, 7]

2) Speaker Segmentation and clustering

Based on the literature review, possible algorithms that can be considered in this step include but not be limited to:

* Bayesian Information Criterion(BIC)
* KL-divergence
* Gaussian Mixtures Models
* I Vector

To make the speaker diarization process performed in a real-time/ left-to-right fashion, necessary modification or adaptions are required for the above algorithms.

# Bibliography

|  |  |
| --- | --- |
| [1] | D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05),* vol. V, pp. 953-956, 2005. |
| [2] | S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on audio, speech, and language processing,* vol. 14, 2006. |
| [3] | X. Anguera Miró, "Robust speaker diarization for meetings," *Universitat Politècnica de Catalunya.,* 2006. |
| [4] | X. Anguera, C. Wooters and J. Hernando, "Acoustic Beamforming for Speaker Diarization of meeting," *IEEE Transactions on Audio, Speech, and Language Processing,* pp. 2011-2022, 2007. |
| [5] | M. Kunešová, Z. Zajíc and V. Radová, "Experiments with Segmentation in an Online Speaker Diarization System," *International Conference on Text, Speech, and Dialogue,* pp. 429-437, 2017. |
| [6] | X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 20, no. 2, pp. 356-370, 2012. |
| [7] | W. Zhu and J. Pelecanos, "Online speaker diarization using adapted i-vector transforms," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5045-5049, 2016. |
| [8] | S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada and S. Makino, "A DOA based speaker diarization system for real meetings," *Hands-Free Speech Communication and Microphone Arrays,* pp. 29-32, 2008. |
| [9] | J. Geiger, F. Wallhoff and G. Rigoll, "GMM-UBM based open-set online speaker diarization," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010. |
| [10] | D. Istrate, C. Fredouille, S. Meignier, L. Besacier and J. Bonastre, "NIST RT’05S evaluation: pre-processing techniques and speaker diarization on multiple microphone meetings," in *International Workshop on Machine Learning for Multimodal Interaction*, Springer, Berlin, 2005, pp. 428-439. |
| [11] | D. A. v. Leeuwen, "The TNO speaker diarization system for NIST RT05s meeting data," in *International Workshop on Machine Learning for Multimodal Interaction*, Springer, Berlin, 2005, pp. 440-449. |
| [12] | T. H. Nguyen, E. S. Chng and H. Li, "Speaker Diarization: An Emerging Research," in *Speech and Audio Processing for Coding, Enhancement and Recognition*, New York, Springer, 2015, pp. 229-277. |
| [13] | "WebRTC," [Online]. Available: https://webrtc.org/. [Accessed 2019]. |
| [14] | J. H. Ko, J. Fromm, M. Philipose, I. Tashev and S. Zarar, "Limiting Numerical Precision of Neural Networks to Achieve Real-Time Voice Activity Detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 2018. |
| [15] | "https://github.com/wiseman/py-webrtcvad," 2019. [Online]. Available: https://github.com/wiseman/py-webrtcvad. [Accessed June 2019]. |
| [16] | P. S. Jadhav, "Classification of Musical Instruments sounds by Using MFCC and Timbral Audio Descriptors," *International Journal on Recent and Innovation Trends in Computing and Communication,* vol. 3, no. 7, pp. 52-56, 2017. |
| [17] | B. McFee, C. Raffe, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th python in science conference*, 2015. |
| [18] | F. Castaldo, D. Colibro, E. Dalmasso, P. Laface and C. Vair, "Stream-based speaker segmentation using speaker factors and eigenvoices," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008. |
| [19] | L. Wan, Q. Wang, A. Papir and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. |
| [20] | X. Anguera and J.-F. Bonastre, "A Novel Speaker Binary Key Derived from Anchor Models," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010. |
| [21] | X. Anguera and J.-F. Bonastre, "Fast speaker diarization based on binary keys," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011. |
| [22] | ICSI Meetings Recorder corpus, 2006. [Online]. Available: http://www1.icsi.berkeley.edu/Speech/mr/. |
| [23] | J. S. Garofolo, M. Michel, V. M. Stanford, E. Tabassi, J. G. Fiscus, C. D. Laprun, N. Pratz and J. Lard, "NIST Meeting Pilot Corpus Speech," 2004. |
| [24] | A. Canavan, D. Graff and G. Zipperlen, "CALLHOME American English Speech," 1997. [Online]. Available: https://catalog.ldc.upenn.edu/ldc97s42. |
| [25] | D. Graff and J. Alabiso, "1996 English Broadcast News Transcripts (HUB4)," 1997. |
| [26] | D. Graff, "An overview of Broadcast News corpora," *Speech Communication,* vol. 37, no. 1-2, pp. 15-26, 2002. |
| [27] | B. Fergani, M. Davy and A. Houacine, "Speaker diarization using one-class support vector machines," *Speech Communication,* pp. 355-365, 2008. |
| [28] | D. Garcia-Romero, D. Snyder, G. Sell, D. Povey and A. McCree, "Speaker diarization using deep neural network embeddings," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 4930-4934, 2017. |
| [29] | D. Liu and F. Kubala, "Online speaker clustering," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP '03,* vol. 1, 2003. |
| [30] | G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," *2014 IEEE Spoken Language Technology Workshop (SLT),* pp. 413-417, 2014. |
| [31] | S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds and J. Glass, "Exploiting intra-conversation variability for speaker diarization," *Twelfth Annual Conference of the International Speech Communication Association,* 2011. |
| [32] | R. Wang, M. Gu, L. Li, M. Xu and T. F. Zheng, "Speaker segmentation using deep speaker vectors for fast speaker change scenarios," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* pp. 5420-5424, 2017. |
| [33] | I. D. Gebru, S. Ba, X. Li and R. Horaud, "Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusio," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 40, no. 5, pp. 1086-1099, 2018. |
| [34] | G. Schwarz, "Estimating the dimension of a model," *The annals of statistics,* vol. 6(2), pp. 461-464, 1978. |
| [35] | X. Anguera, C. Wooters, B. Peskin and M. Aguiló, "Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System," *International Workshop on Machine Learning for Multimodal Interaction,* pp. 402-414, 2005. |
| [36] | M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin and S. Meignier, "An Open-source State-of-the-art Toolbox for Broadcast News Diarization," *Interspeech,* 2013. |
| [37] | G. D. P. G. E. K. T. M. S. M. Mickael Rouvier, "An Open-source State-of-the-art Toolbox for Broadcast News Diarization," *Interspeech,* 2013. |
| [38] | S. S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," *Proc. DARPA broadcast news transcription and understanding workshop,* vol. 8, pp. 127-132, 1998. |
| [39] | Y. Huang, O. Vinyals, G. Friedland, C. Müller, N. Mirghafori and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU),* pp. 693-698, 2008. |
| [40] | M. A. Siegler, U. Jain, B. Raj and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," *Proc. DARPA speech recognition workshop,* 1997. |
| [41] | H. Gish, M. H. Siu and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," *1991 International Conference on Acoustics, Speech, and Signal Processing,* pp. 873-876, 1991. |
| [42] | X. Anguera, C. Wooters, B. Peskin and M. Aguiló, "Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System," *International Workshop on Machine Learning for Multimodal Interaction,* pp. 402-414, 2005. |
| [43] | "Librosa," 2018. [Online]. Available: https://librosa.github.io/librosa/. [Accessed Jun 2019]. |