**Real-time Speaker Diarization System**

**Dissertation Interim Report**

by Pan Hao

Presented to the Hong Kong University

Department of Computer Science

In Partial Fulfillment of the Requirements for the Degree of

Master of Science in Computer Science

Dr. C.L. Yip, BEng, PhD (Supervisor)

May 2019

1. **Introduction**

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 (Reynolds & Torres-Carrasquillo, 2005) . 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.

1. **Literature Review**

The traditional three primary application domains of speaker diarization is broadcast news, recorded meetings and telephone conversation (Tranter & Reynolds, 2006). 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. Anguera Miró (2006) 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, Anguera, Wooters and Hernando (2007) 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 (Kunešová, Zajíc, & Radová, 2017). 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 (Tranter & Reynolds, 2006; Miro, et al., 2012). One popular segmentation algorithms is the use of Bayesian information criterion (BIC) firstly introduced by Schwarz (1978) and firstly used in speaker segmentation by Chen and Gopalakrishnan (1998). Many state-of-art systems incorporate BIC as a segmentation metric in the following (Anguera, Wooters, Peskin, & Aguiló, 2005; Mickael Rouvier, 2013). As BIC approach is computationally intensive, several works (e.g. (Huang, et al., 2008)) propose modification or other technologies used with BIC to speed up the process. Some common alternative segmentation approaches include Generalized Likelihood Ratio (GLR) (Gish, Siu, & Rohlicek, 1991) and Kullback–Leibler (KL) divergence (Siegler, Jain, Raj, & Stern, 1997). Some recent papers propose advanced technology like deep neural network (DNN) to find speaker change points (Wang, Gu, Li, Xu, & Zheng, 2017).

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 (2014) 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 (Zhu & Pelecanos, 2016) 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 (Anguera Miró, 2006) and there is limited work on online speaker diarization. Araki, Fujimoto, Ishizuka, Sawada and Makino (2008) present 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. Geiger, Wallhoff and Rigoll (2010) 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. Zhu and Pelecanos (2016) 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.

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 (ICSI Meetings Recorder corpus, 2006)
* NIST Meeting Pilot Corpus Speech (Garofolo, et al., 2004)

For telephone conversation, sample databases include:

* LDC CALLHOME English corpus (Canavan, Graff, & Zipperlen, 1997) and CALLHOME corpus in other languages are also available.

For broadcast news, sample databases include:

* LDC 1996 radio broadcast news database (HUB4) (Graff & Alabiso, 1997) and an overview of broadcast news corpora is conducted by Graff (2002).
  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) (Miro, et al., 2012). 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) (Anguera Miró, 2006).

In this dissertation project, the focus will be put on the MFCCs which are frequently used in online speaker diarization system (Geiger, Wallhoff, & Rigoll, 2010; Zhu & Pelecanos, 2016)

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

Anguera Miró, X. (2006). Robust speaker diarization for meetings. *Universitat Politècnica de Catalunya.*

Anguera, X., Wooters, C., & Hernando, J. (2007). Acoustic Beamforming for Speaker Diarization of meeting. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011-2022.

Anguera, X., Wooters, C., Peskin, B., & Aguiló, M. (2005). Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System. *International Workshop on Machine Learning for Multimodal Interaction*, 402-414.

Anguera, X., Wooters, C., Peskin, B., & Aguiló, M. (2005). Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System. *International Workshop on Machine Learning for Multimodal Interaction*, 402-414.

Araki, S., Fujimoto, M., Ishizuka, K., Sawada, H., & Makino, S. (2008). A DOA based speaker diarization system for real meetings. *Hands-Free Speech Communication and Microphone Arrays*, 29-32.

Canavan, A., Graff, D., & Zipperlen, G. (1997). *CALLHOME American English Speech*. Retrieved from https://catalog.ldc.upenn.edu/ldc97s42

Chen, S. S., & Gopalakrishnan, P. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. *Proc. DARPA broadcast news transcription and understanding workshop, 8*, 127-132.

Fergani, B., Davy, M., & Houacine, A. (2008). Speaker diarization using one-class support vector machines. *Speech Communication*, 355-365.

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

Garofolo, J. S., Michel, M., Stanford, V. M., Tabassi, E., Fiscus, J. G., Laprun, C. D., . . . Lard, J. (2004). NIST Meeting Pilot Corpus Speech. Retrieved from https://catalog.ldc.upenn.edu/LDC2004S09

Gebru, I. D., Ba, S., Li, X., & Horaud, R. (2018). Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusio. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(5), 1086-1099.

Geiger, J., Wallhoff, F., & Rigoll, G. (2010). GMM-UBM based open-set online speaker diarization. *Eleventh Annual Conference of the International Speech Communication Association.*

Gish, H., Siu, M. H., & Rohlicek, R. (1991). Segregation of speakers for speech recognition and speaker identification. *1991 International Conference on Acoustics, Speech, and Signal Processing*, 873-876.

Graff, D. (2002). An overview of Broadcast News corpora. *Speech Communication, 37*(1-2), 15-26.

Graff, D., & Alabiso, J. (1997). 1996 English Broadcast News Transcripts (HUB4). Retrieved from https://catalog.ldc.upenn.edu/LDC97T22

Huang, Y., Vinyals, O., Friedland, G., Müller, C., Mirghafori, N., & Wooters, C. (2008). A fast-match approach for robust, faster than real-time speaker diarization. *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 693-698.

ICSI Meetings Recorder corpus. (2006). Retrieved from http://www1.icsi.berkeley.edu/Speech/mr/

Kunešová, M., Zajíc, Z., & Radová, V. (2017). Experiments with Segmentation in an Online Speaker Diarization System. *International Conference on Text, Speech, and Dialogue*, 429-437.

Liu, D., & Kubala, F. (2003). Online speaker clustering. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP '03, 1*.

Mickael Rouvier, G. D. (2013). An Open-source State-of-the-art Toolbox for Broadcast News Diarization. *Interspeech*.

Miro, X. A., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(2), 356-370.

Reynolds, D. A., & Torres-Carrasquillo, P. (2005). Approaches and Applications of Audio Diarization. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05), V*, 953-956.

Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., & Meignier, S. (2013). An Open-source State-of-the-art Toolbox for Broadcast News Diarization. *Interspeech*.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics, 6(2)*, 461-464.

Sell, G., & Garcia-Romero, D. (2014). Speaker diarization with plda i-vector scoring and unsupervised calibration. *2014 IEEE Spoken Language Technology Workshop (SLT)*, 413-417.

Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D., & Glass, J. (2011). Exploiting intra-conversation variability for speaker diarization. *Twelfth Annual Conference of the International Speech Communication Association*.

Siegler, M. A., Jain, U., Raj, B., & Stern, R. M. (1997). Automatic segmentation, classification and clustering of broadcast news audio. *Proc. DARPA speech recognition workshop*.

Tranter, S. E., & Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing, 14*.

Wang, R., Gu, M., Li, L., Xu, M., & Zheng, T. F. (2017). Speaker segmentation using deep speaker vectors for fast speaker change scenarios. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5420-5424.

Zhu, W., & Pelecanos, J. (2016). Online speaker diarization using adapted i-vector transforms. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5045-5049.