# Speaker Diarization: Its Developments, Applications, And Challenges

Hernawan Sulistyanto, *Member, IEEE*
School of Computer Science and Software Engineering
University of Wollongong
NSW, Australia
Email : hs984@uowmail.edu.au

Informatics Engineering
Muhammadiyah University of Surakarta
Central Java, Indonesia
Email : hnstyanto@yahoo.com

***Abstract*—Multimedia computing is an emerging research area with a wide range of possibilities for various types of digital media. For instance television, movie, audio broadcast, and meeting recordings can be analysed to derive semantic value. Although significant progress has been achieved in recent years, speaker diarization continues to be an active topic in speech research. In this paper, a speaker diarization system would be reviewed and then continued by description of current development and application of diarization system in the broadcast news, meeting and telephone channel. Some trends and challenges arising in the speaker diarization area are addressed in this review, such as presence of overlapped speech and various characteristic of the input audio stream. Particularly, performance improvement has become an ongoing goal in speaker diarization research. One of important factors is how far speaker diarization system has been implemented in many widely area to assist other application systems.**

***Keywords : component; speaker diarization, broadcast news, meeting, telephone channel, overlapped speech***

## I. INTRODUCTION

One of multimedia components that currently intensively researched by research groups is audio stream. Although most researches concentrate on jointly visual and audio channel analysis, individual audio channel analysing separately is essential and contributable to multimedia content analysis. Diarization has recently received much attention. Audio diarization could be widely defined in various ways, but a common core of the definition remains, audio diarization is the process of annotating audio stream with information that attributes temporal regions of energy to specific source. When related to speaker, this is a task of automatically partitioning an input audio stream into homogeneous segments and assigning these segments to specific source. These sources generally include particular speaker, music or background noise, such is called 'cocktail party problem'.

Recently, speaker diarization has been an exciting topic in the speaker recognition research community. The main problem would be solved by speaker diarization is to determine the number of people involved in a conversation of either in the meeting, broadcast news, or telephone. Moreover, the speaker diarization serves the problem solving for indexing and retrieval-based systems.

Speaker diarization has consumed attention in the recent years. Some researches dealing with speaker diarization development have been introduced by the authors. Hung, et.al., in [1] proposed method to determine the dominance applying diarization. Multi modal speaker diarization was conducted by Noulas, et.al., in [2]. A linguistic influence on method of speaker diarization was also repoted by Bozonnet, et.al., in [3]. Speaker diarization was also investigated in web audio file application by Clement, at.al., in[4]. In [5], Bozonnet, at.al., presented a multimodal approach to speaker diarization of TV show data. Speaker dfiarization in meeting using intensity channel contribution was handled by Barra-Chicote, et.al., in [6]. Hence, it is clearly shown that SD presents to serve other systems.

Existence of speaker diarization system among other systems has contributed in the signal processing area, especially for advanced speaker recognition. Some application sides of speaker diarization system would be described in the following sections of this paper. The most important of current speaker diarization system is at paying attention on some challenges arising, such as audio stream that contains overlapped speech, and heterogenity of input audio. Our contribution in this paper is at addressing the trends of diarization research, exhibiting the application and awareness from advance challenges in this area.

Overall, the rest of paper is organised into these sections. Current speaker diarization architecture would be described in section 2. In section 3, it is going to present direction of speaker diarization developments and some current application. Some result related to speaker diarization application would be introduced in section 4.

Finally, section 5 releases some valuable conclusion in the research area of speaker diarization.

## II. CURRENT SPEAKER DIARIZATION SYSTEM

Some terms have recently defined the speaker diarization. Generally, speaker diarization is a task to determine the number of speaker who is active and their utterance duration in an observed audio stream. Speaker diarization is different compared to two existing popular systems in the recognition are, i.e., speaker identification and speech recognition. Speaker identification is at determination the speaker identity (who is speaking) and the speech recognition is to get the words spoken by the speaker (what spoken). Whereas, speaker diarization is to solve the problem "who spoke when". The main goal of speaker diarization is at extracting segments of speech and associating them with the correct speaker. The speaker labels derived by diarization process are relative to audio recording and they indicate which audio segments were spoken by the same speaker. However, they do not denote the true identity of the speaker. Depiction for generally speaker diarization task shown below:

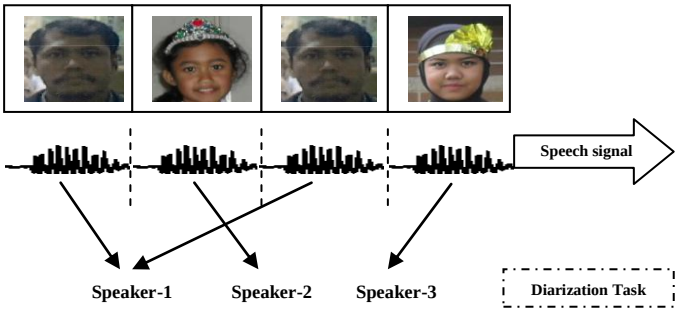


FIGURE 1. OVERVIEW OF SPEAKER DIARIZATION TASK.

Common speaker diarization system is generally depicted on the following Figure 2 and 3. Speaker diarization system is composed by two main stages, namely segmentation and clustering stages. These stages could be modified depend on approach employed and aim of its application. Input of system is entire audio stream containing speech signal. This speech signal could be refined from some media either direct speech or recording. Speaker segmentation aims to partition the speech signal into homogeneous segments. This stage is commonly initialized by a pre-processing step to extract speech signal into features companied by windowing. In this segmentation, there is speaker change detection which is to identify the speakers turning in a conversation or dialog scene. Obtained segments are then grouped into certain class according to specific speakers. This process occurred in the speaker clustering stage. Result of segments grouped into clusters is then used to construct the speaker models. In the current speaker diarization system, one cluster would be assigned to represent one speaker. Based on the number of unique clusters obtained, it will be able to determine the number of speakers who present in the observed speech signal. Finally, clusters modelling the speaker are labelled as speaker-1, speaker-2, etc.

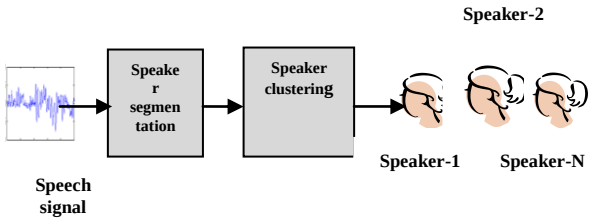


FIGURE 2. DIAGRAM OF DIARIZATION FRAMEWORK.

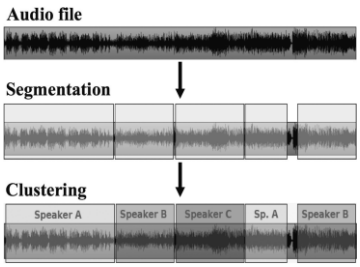


FIGURE 3. SPEAKER DIARIZATION MECHANISM.

## III. METHOD AND EVALUATION

### A. *Approaches and Tecgniques*

Speaker diarization research has been performed with many approaches and techniques. All of them are characterized in features extraction, speaker segmentation, and speaker clustering. Speaker diarization is a speech-processing technique. There are many well-known acoustic features have been employed in the audio signal extraction. Mel Frequency Cepstral Coefficients (MFCCs) is the most common acoustic feature choice for speaker diarization. Speaker segmentation is commonly performed by using energy-based, model-based, or measure-based segmentation. In the energy-based segmentation, the analysis is based on the acoustic energy of the audio stream. The common choice for model-based approaches is applying Hidden Markov Model (HMM) and Gaussian Mixture Models (GMMs). GMMs are frequently chosen because they are universal density approximators, i.e., they can model an arbitrary probability distribution function over the data. Measure-based approaches measure the difference between two consecutive segments of the audio stream, which is usually referred to as distance between the

two segments. Speaker clustering approaches is categorized into hierarchical-based and model-based clustering. The hierarchical clustering is an intuitive method to cluster the audio segments. Initially, the distance between each pair of speech segments is computed, using a user defined distance measure which assigns smaller distances to acoustically similar segments. The model-based clustering is typically performed in parallel with the segmentation. Multiple passes over the audio stream create the optimal segmentation, cluster the data, adjust the cluster model parameters and re-segment the audio track. This procedure is iterated until convergence or until a stopping criterion is met.

### *B. Evaluation of Speaker Diarization Performance*

Process done by speaker diarization system has a number of approaches and methods, which have been investigated to warranty the diarization accuracy. Speaker diarization performance is evaluated by a force alignment between diarizaton hypothesize with a segments reference produced by manually labelling. This performance assessment involves a number of parameters consist of missed speech (MS), false alarm (FA), and speaker error (SE). The sum of these assessment parameters is represented with Diaization Error Rate (DER). DER mathematically is defined as:

$$DER = MS + FA + SE \quad (1)$$

Where MS is occurrence when there is not speech in hypothesize but present in reference or other word it does not detect the presence of speech. On the contrary, if speech is detected in hypothesize on which actually there is no speech in reference, is called FA. SE is defined as fault in speaker detection. Mapping of segments between hypothesize and reference is shown below.

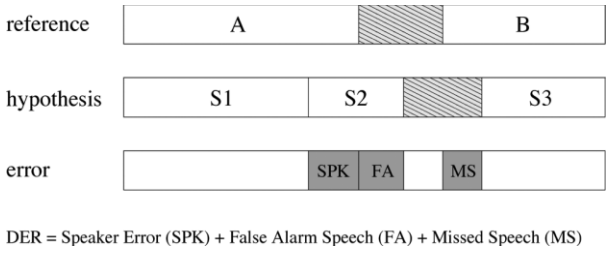


FIGURE 4. SEGMENT ALLIGMENT TO EVALUATE DIARIZATION PERFORMANCE.

## IV. TREND FOR DEVELOPMENTS AND APPLICATION OF SPEAKER DIARIZATION SYSTEM

Speaker diarization system has many utilities in the audio-based application. For instance, application for audio and speaker indexing, content structuring, audio information retrieval, speaker verification (in the presence of multiple-speakers), speech-to-text transcription, and some are of video processing.

There are generally three primary application domain for speaker diarization research and development, namely broadcast news audio, recorded meeting and telephone conversation. These domain data differ in the quality of the recordings (bandwidth, microphone, noise) the amount and type of non-speech sources, the number of speakers, and the style and structure of the speech (e.g., scripted, duration, and sequencing of speaker turns). In [7][8][9], speaker diarization system is implemented in broadcast news. In this application, diarization system is utilized to label segments into commercial break, speech and non-speech using multistage diarization. Most speaker diarization system for broadcast news data has similar general architecture. The signal is chopped into homogeneous segments. The segments boundaries are located by finding acoustic changes in the signal and each segment is expected to contain speech from only one speaker. The resulted segments are then clustered so that each cluster corresponds to one speaker.

Speaker diarization is applied in meeting [10][11][12][13][14][15] to determine the number of participant speaking in a meeting room. Hence, it has also been advanced to monitor the most active participant. In the advance development, speaker diarzation has also been applied to consider the dominance of active speakers. In this application commonly need a set of microphone used by each meeting's member. Application in telephone conversation was introduced in [16]. Summary of DER derived by diarization applications are conveyed below.

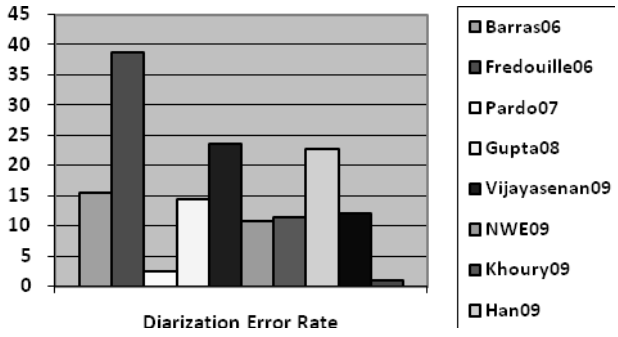


FIGURE 5. DER RESULTS DERIVED BY THE EXPERIMENTS.

Understanding of table presented above, it needs to be under lined that obtained DER results are neither a absolute value nor a depiction which states a DER value result is better than another one. It is caused aforementioned DER value is obtained by experiments using various technique, and employing different feature and datasets or corpus.

## V. CHALLENGES IN SPEAKER DIARIZATION

Research in speaker diarization has emerged a number of challenges in its developments. In the first one is related to dataset/corpus implemented for the speaker diarization system. Broadcast News (BN) speech is usually acquired using boom or lapel microphones with some recordings being made in the studio and others in the field. Conversely, meetings are usually recorded using desktop or far-field microphones (single microphones or microphone arrays) which are more convenient for users than head-mounted or lapel microphones. As a result the signal-to noise ratio is generally better for BN data than meeting recordings. Additionally, differences between meeting room configurations and microphone placement lead to variations in recording quality, including background noise, reverberation and variable speech levels (depending on the distance between speakers and microphones). BN speech is also often read or at least prepared in advance while meeting speech tends to be more spontaneous in nature and contains more overlapping speech. Although BN recordings can contain speech that is overlapped with music, laughter, or applause (far less common for conference meeting data), in general, the detection of acoustic events and speakers tends to be more challenging for conference meeting data than for BN data. Finally, the number of speakers is usually larger in BN but speaker turns occur less frequently than they do in conference meeting data, resulting in BN having a longer average speaker turn length. Movie dataset is rarely used to currently examine the diarization system performance. A few researches in other works have employed movie dataset so far. Hence, it is clearly seen that movie dataset still widely open a new chance to be investigated.

TABLE I. IMPROVED DER BY USING THE OVERLAPPED SPEECH DETECTION IN SPEAKER DIARIZATION

**Error! Not a valid link.**

The second challenge is concerned to existence of overlapped speech. Speakers in a conversation or a dialogue are distinguishable in two main ways; they may speak alternately and or simultaneously. In spontaneous and natural conversation, two or more people speaking simultaneously are a common occurrence. The simultaneous speech from more than one speaker is referred as overlapped speech. Overlapped speech presents a new challenge to automatic systems that process the audio data. An example is a speaker diarization system. In a few researches has been shown that presence of overlapped speech contributes a significant number of errors in most state-of-the-art speaker diarizaton system. These errors would generally appear in the form of missed speech and speaker errors. It needs to consider and aware against the existence of overlapped speech in the further system development.

In other word, diarization system that involves overlapped speech detection will have better performance compared to diarization system that ignores presence of overlapped speech. It has been proven by experiment results presented Table above.

## VI. CONCLUSION

Speaker diarization system aims to serve the speaker indexing task in a given audio stream. Some applications obviously contributed by speaker diarization system have been presented.

The current new challenges have also addressed to be a starting point for further diarization developments. It means that the advance speaker diarization will need more attention at ability to process the input audio stream in real environment, such as the likelihood of usage of party recording in real-time diarization and even on-line dataset. Furthermore, advance speaker diarization should be able to handle presence of overlapped speech on which the occurrence of overlapping speech almost regularly presents in natural conversation. In case of overlapped speech handling, it is still widely opened the space to researches, mainly in context speech segregation for overlapped speech handling, in which current approach to handle overlapping speech is with removal the segments that contain overlapped speech.

## ACKNOWLEDGMENT

This work has being supported by the General Directorate for Higher Educational Degree with BATCH IIIA of National Educational Minister of Indonesia where author comes from.

## REFERENCES

[1] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating Dominance in Multi-Party Meetings Using Speaker Diarization", IEEE Trans. On Audio Speech and Language Processing, v.19, no.4, pp847-860, May 2011.

[2] A. Noulas, G. Englebienne, and B. J.A. Krose, "Multimodal Speaker Diarization", IEEE Trans. On Pattern Analysis and Machine Intelligent, pp1-35, 2011.

[3] S. Bozonnet, D. Wang, N.Evans and R. Troncy, "Lingusitic Influences on Bottom-up and Top-down for Speaker Diarization", In Proc. of ICASSP, 2011, pp4424-4427.

[4] P. Clement, T. Bazillon, C. Fredouille, "Speaker diarization of Heterogeneous Web Video Files: A Preliminary study", In Proc of ICASSP, 2011, pp4432-4435.

[5] S. Bozonnet, F. Vallet, N. Evans, S. Essid, G. Richard and J. Carrive, "A multi modal approach to Initialisation for Top-down Speaker Diarization of Television Shows", French Institut National de l'Audiovisuel (INA), 2011.

[6] R. Barra-Chicote, J. M. Pardo, J. Ferreiros,, and J. M. Montero, "Diarization Based on Intensity Channel Contribution", IEEE Trans. On ASLP, v.19, no.4 pp754-761, May, 2011.

[7] C. Barras, X. Zhu, S. Meignier, and J. Gauvain, "Multistage Speaker Diarization of Broadcast News", IEEE Trans. ASLP, v.14, no.5, pp1505-1512, September, 2006.

[8] V. Gupta, G. Boulianne, P. Kenny, P. Ouellet, and P. Dumouchel, "Sepaker Diarizatiob of French Broadcast News", In IEEE Proc. of ICASSP, 2008, pp4365-4368.

[9] O. Yılmaz and M. Saraclar, "Diarization for Turkish Broadcast News Transcription", IEEE 19th Signal Processing and Communications Applications Conference (SIU 2011) Audio, 2011, pp379-382.

[10] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker Diarization for Multiple-Distant-Microphone Meetings Using Several Sources of Information", IEEE Trans. On Computers, VOL. 56, NO. 9, pp1212-1224, September 2007.

[11] D. Vijayasenan, F. Valente, H. Bourlard, "Mutual Information Based Channel Selection for Speaker Diarization of Meeting Data", In IEEE Proc of ICASSP, 2009, pp 4065-4069.

[12] T. L. Nwe, H. Sun, H. Li and S. Rahardja, "Speaker Diarization in Meeting Audio", In IEEE Proc. of ICASSP, 2009, pp4073-4073.

[13] E. El-Khoury, C. Sénac, and . J. Pinquier, "Improved Speaker Diarization System for Meetings", In IEEE Proc of ICASSP, 2009, pp4097-4100.

[14] H. Sun, B. Ma, S.Z. K. Khine and H. Li, "Speaker Diarization System for RT)& and RT 09 Meeting Room Audio", In Proc of ICASSP, 2010, pp4982-4985.

[15] G. Friedland, J. Chong, and A. Janin, "Parallelizing Speaker-Attributed Speech Recognition for Meeting Browsing", IEEE International Symposium on Multimedia, 2010, pp121-128.

[16] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of Telephone Conversations Using Factor Analysis", IEEE Journal of Selected Topics in Signal Processing, vol4, no6, pp1059-1070, December, 2010.

[17] O. Ben-Harus, H. Guterman, and I. Lapidot, "Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization", In Proc. of MLSP, 2009, pp1-6.

[18] K. Boakye, B. Trueba-Horneo, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings", In Proc. of IEEE International Conference on Acoustics, Speech and signal Processing (ICASSP), pp4353-4356, 2008.

[19] K. Boakye, O. Vinyals, and G. Friedland. "Two's a crowd: Imprpvement speaker diarization by automatically identifying and excluding overlapped speech", In Interspeech, pp32-35, Brisbane, Australia, September 2008.

[20] M. Huijbregts, D. V. Leeuwen, and F. D. Jong, "Speech overlap detection in a two-pass speaker diarization system", In Interspeech, pp1063-1066, 2009.

[21] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization", In Proc. of ASRU, pp 683-686, 2007

[22] M. Zelenak and J. Hernando, "On the improvement of speaker diarization by detecting overlapped speech", In VI Jornadas en Tecnologa del Habla and II Iberian SLTech Workshop, pp153-156, 2010.