International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 5 | ISSN : 2456-3307

One-Decade Survey on Speaker Diarization for Telephone and Meeting Speech

Ajit Das*[1], Utpal Bhattacharjee[2], Dipak Kr. Mitra[3]

[*1] Department of CST, Bodoland University, Kokrajhar, Assam, India
adas0078@rediffmail.com
[2] Department of IT, Rajiv Gandhi University, Arunachal Pradesh, India
utpal.bhattacharjee@rgu.ac.in
[3] Department of Mathematical Sciences, Bodoland University, Kokrajhar, Assam, India
dkrmitra@gmail.com

ABSTRACT

Finding speaker turns and identifying the speakers is known as speaker diarizationi.e speakerdiarization effectively answer the question ‘who speak and when’. In other words its task is to determine the speaker turns in an audio or video recording which contents unknown speech and unknown number of speakers. Over recent years this domains have received most research attention within the speaker diarization community. It is mainly used in many applications related to audio processing such as information retrieval from telephone conversation, meeting speech, broadcast news etc. In this paper, our aim is to review the current state-of-the-art, focusing on research developed since beginning of diarization that relates to Speaker Diarization for telephone and meeting speech.

Keywords: Speaker Diarization, Meeting Speech,Telephone Speech, Segmentation and Clustering.

I. INTRODUCTION

Speaker diarization can be summarized as the “Who spoke when” problem. Detection of speech and non-speech can be considered as a basic segmentation system [1]. Comprehensive segmentation systems can include gender classification and organization of the input source and detection of speech, speaker turn points and narrow band speech. Normally, two types of segmentation systems are considered: agglomerative-based baseline system and variational Bayesian diarization [2]. In addition, Maximum Likelihood Speech Source Separation is proposed for detection of simultaneous speech segments and identities of speakers in this segment.

Front end Processing:

First stage of speaker recognition or segmentation is speech parameterization of the input signal on short term spectral content. In the process of speech parameterization, input signal is converted into a vector consisting of features. The purpose of conversion from input signal to a new representation is to have compact, less redundant and more suitable representation of speech signal for statistical modelling and scoring calculations. Mel Frequency Cepstral Coefficients (MFCC) [3], Perceptual Linear Predictive Coding (PLP) [4] and Linear Predictive Coding (LPC) [5] are proposed for speech parameterization methods. Most of the systems use MFCC representation is used in our studies because Mel-scale and logarithmically spaced filters that is utilized in MFCC are good approximation of human auditory system.

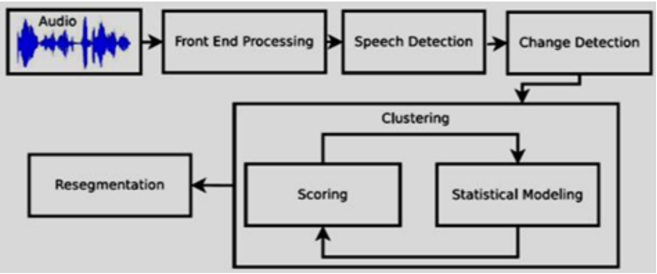

Figure 1. Block Diagram of Speaker Diarization

CSEIT1725229| Received: 16 Oct 2017 | Accepted: 31 Oct 2017 | September-October-2017[(2)5: 990-994]

II. SPEAKER DIARIZATION FOR MEETING SPEECH

Speaker Diarization for meeting speech is a challenging topic for the researchers. Here I have mentioned few researches done by various researchers in meeting speech:

Authors Elie El-Khoury, Christine Senac and JulienPinquier in their paper "Improved Speaker Diarization System For Meetings" tried to decrease the error rate of speaker diarization for meting speech by using new iterative scheme. Here, authors proposed new bidirectional source segment based on GLR/BIC method. They reviewed early BIC clustering method and proposed a new unsupervised post processing is added to increase cluster purity. In this experiment they proved 40% improvement compared to speaker diarization system [6]. Author DeepuVijayasenan in his paper "An Information Theoretic Approach to Speaker Diarization of Meeting Data" described speaker diarization system based on an information theoretic framework. Here it is discussed issues related to speaker diarization using this information theoretic framework such as the criteria for inferring the number of speakers, the trade off between quality and compression achieved by the diarization system, and the algorithms for optimizing the objective function. The problem is formulated according to the Information Bottleneck (IB) principle. This solves the problem of choosing the distance between speech segments, which becomes the Jensen–Shannon divergence as it arises from the IB objective function optimization discussed issues related to speaker diarization using this information theoretic framework such as the criteria for inferring the number of speakers, the trade off between quality and compression achieved by the diarization system, and the algorithms for optimizing the objective function. They compared it with NIST RT06 (Rich Transcription) data set for speaker diarization of meetings. The IB-based system achieves a diarization error rate of 23.2% compared to 23.6% for the baseline system. This approach being mainly based on nonparametric clustering, it runs significantly faster than the baseline HMM/GMM based system, resulting in faster-than-real-time diarization[7]. Authors Fabio Valente, PetrMotlicek and DeepuVijayasenan in their paper "Variational Bayesian Speaker Diarization Of Meeting Recordings" investigates the use of the Variational Bayesian (VB) framework for speaker diarization of meetings data VB learning aims at maximizing a bound, known as Free Energy, on the model marginal likelihood and allows joint model learning and model selection according to the same objective function. While the BIC is valid only in the asymptotic limit, the Free Energy is always a valid bound. In this paper, the authors proposes the use of Free Energy as objective function in speaker diarization. It can be used to select dynamically without any supervision or tuning, elements that typically affect the diarization performance i.e. the inferred number of speakers, the size of the GMM and the initialization. The proposed approach is compared with a conventional state-of-the-art system on the RT06 evaluation data for meeting recordings diarization and shows an improvement of 8.4% relative in terms of speaker error [8]. Authors Wei Li Yanxiong Li and Qianhua He in their paper "Estimating Key Speaker in Meeting Speech Basedon Multiple Features Optimization" proposed to estimate key speaker in meeting speech based on multiplefeatures optimization. First, each feature is defined and their differences between key speaker and other speakers are analyzed. Then, a decision function of multiple feature weighting is generated for estimating key speaker in meeting speech, and the genetic algorithm is used to optimize these coefficients of feature weighting. The methods are evaluated on three different meeting speech datasets. Experimental results show that the proposed optimization method obtains average accuracy of 93.3% for estimating key speaker, and gains average accuracy improvement by 9.7% [9]. Authors Hayley Hung , Yan Huang, Gerald Friedland and Daniel Gatica-Perez in their paper "Estimating Dominance in Multi-Party Meetings Using Speaker Diarization" investigate the task of automatically measuring dominance in small group meetings when only a single audio source is available. For these tasks they use speaker segmentations generated by our automated faster than real-time speaker diarization algorithm, where the number of speakers unknown. From user-annotated data, they analyse how the inherent variability of the annotations affects the performance of our dominance estimation method. They primarily focus on examining of how the performance of the speaker diarization and our dominance tasks vary under different experimental conditions and computationally efficient strategies, and how this would impact on a practical implementation of such a system. Despite the use of a state-of-the-art speaker diarization algorithm, speaker segments can be noisy. On conducting experiments on almost 5 hours of audio-visual meeting data, their results show that the dominance estimation is robust to increasing diarization noise [10].

Authors Qiao Li, Qing Fan, Yunpeng Xiao and Weiping Ye in their paper ”A Comparable Study on PNCC in Speaker Diarization for Meetings” proposed a new approach i.e Power Normalized Cepstrum Coefficients (PNCC) achieves impressive improvement in noisy speech recognition compare to MFCC. It consequently expects a proof for speaker diarization use. In this paper, PNCC is evaluated against MFCC in a meeting domain speaker diarization system. The Diarization Error Rate (DER) shows no positive results with PNCC. This is possibly because of their inhibition in high frequency spectrum which is believed to represents

[14]. the characteristics of human's voice[11]. Authors SreeHarshaYella, HervéBourlard in their paper “Overlapping Speech Detection Using Long-Term Conversational Features for Speaker Diarization in Meeting Room Conversations” mentioned a challenge how to deal with overlapped speech that has been identified as one of the main source of errors in speaker diarization. Overlapped occurrence is correlated with various conversational features such as speech, silence patterns and speaker turn changes. Here they use features capturing this higher level information from structure of a conversation such as silence and speaker change statistics to improve acoustic feature based classifier of overlapping and single-speaker speech classes. The silence and speaker change statistics are computed over a long-term window (around 3-4 seconds) and are used to predict the probability of overlap in the window. These estimates are then incorporated into a acoustic feature based classifier as prior probabilities of the classes. Experiments conducted on three corpora (AMI, NIST-RT and ICSI) have shown that the proposed method improves the performance of acoustic feature based overlap detector on all the corpora. They also reveal that the model based on long-term conversational features used to estimate probability of overlap which is learned from AMI corpus generalizes to meetings from other corpora (NIST-RT and ICSI) [12]. Authors Giovanni Soldi, Christophe Beaugeant and Nicholas Evans in their paper “Adaptive And Online Speaker Diarization For Meeting Data” have been mention the importance of real-time diarization, stemming from the increasing popularity of powerful, mobile smart devices. While a small number of such systems have been reported, truly online diarization systems for challenging and highly spontaneous meeting data are lacking. This paper reports a work to develop an adaptive and online diarization system using the NIST Rich Transcription meetings corpora [13]. Authors SreeHarshaYella and HerveBourlard in their paper “Information Bottleneck Based Speaker Diarization of Meetings Using Non-Speech as side Information” mentioned that the performance degradation in speaker diarization system caused by Background noise and errors in speech/non-speech detection. In a typical speaker diarization system, non-speech segments are excluded prior to unsupervised clustering. In their current study, they tried to exploit the information present in the non-speech segments of a recording to improve the output of the speaker diarization system based on information bottleneck framework. This is achieved by providing information from non-speech segments as side (irrelevant) information-to-information bottleneck based clustering. Experiments on meeting recordings from RT 06, 07, 09, evaluation sets have shown that the proposed method decreases the diarization error rate by around 18% relative to the baseline speaker diarizationsystem based on information bottleneck framework. Comparison with a state of the art system based on HMM/GMM framework shows that the proposed method significantly decreases the gap in performance between the information bottleneck system and HMM/GMM[14]. Authors Xavier Anguera, Chuck Wooters and Javier Hernando in their paper “Acoustic Beamforming for Speaker Diarization of Meetings” mentioned about the use of microphone in speaker diarization for meetingspeech. Instead of this they tried to establish the use of classic acoustic beamforming techniques together with several novel algorithms to create a complete frontend for speaker diarization in the meeting room domain. New techniques we are presenting include blind reference-channel selection, two-step time delay of arrival (TDOA) Viterbi post processing, and a dynamic output signal weighting algorithm, together with using such TDOA values in the diarization to complement the acoustic information. Tests on speaker diarization show a 25% relative improvement on the test set

compared to using a single most centrally located microphone [15].

III. SPEAKER DIARIZATION FOR TELEPHONE SPEECH

Authors ItshakLapidot, Jean-Francois Bonastre and SamyBengio in their paper " Telephone Conversation Speaker Diarization Using Mealy-HMMs" have shown the use of Mealy- HMMs for telephone conversation speaker diarization task. When Hidden Markov Models (HMMs) were first introduced, two competing representation models were proposed, the Moore model, with separate emission and transition distributions, which is commonly used in speech technologies, and the Mealy model, with a single emission-transition distribution. Since then the literature has mostly focused on the Moore model. In this paper, we would like to show the use of Mealy-HMMs for telephone conversation speaker diarization task. We present the Viterbi training and decoding for Mealy-HMMs and show that it yields similar performance compared to Moore-HMMs with a fewer number of parameters [16]. Authors Simon Bozonnet, RavichanderVipperla and Nicholas Evans in their paper "Phone Adaptive Training for Speaker Diarization" has mention that linguistic content of a speech signal is a source of unwanted variation which can degrade speaker diarization performance. Here to apply a new approach that is called Phone Adaptive Training (PAT), is analogous to speaker adaptive training used in automatic speech recognition. They report an oracle experiment which shows that PAT has the potential to deliver a 33% relative improvement in the diarization error rate of our baseline system [17]. Authors RongZheng, Ce Zhang, Shanshan Zhang and Bo Xu in their paper "Variational Bayes Based I-Vector For Speaker Diarization Of Telephone Conversations" investigated the variational Bayes based I-vector method for speaker diarization of telephone conversations. The motivation of the proposed algorithm is to utilize variational Bayesian framework and exploit potential channel effect of total variability modelling for diarization of conversation side. Other three well-known techniques are compared as follows: K-means clustering for eigen voices and I-vector speaker diarization, and variational Bayes applied to eigen voices. Performance evaluations are conducted on the summed-channel telephone data from the 2008 NIST speaker recognition evaluation. The paper discusses how the performance is influenced by different modules[18].
Authors HoumanGhaemmaghami, David Dean and SridhaSridharan in their paper "A Speaker Rediarization Scheme for Improving Diarization in Large Two-Speaker Telephone Datasets" proposed a approach speaker rediarization in an iterative manner. Their aim was to show the information obtained through the first pass of speaker diarizationcan be reused to refine and improve the original diarization results and they demonstrate the practical application of their rediarization algorithm using a large archive of two-speaker telephone conversation recordings. They use the NIST 2008 SRE summed telephone corpora for evaluating our speaker rediarization system. This corpus contains recurring speaker identities across independent recording sessions that need to be linked across the entire corpus. They show that their speaker rediarization scheme can take advantage of inter-session speaker information, linked in the initial diarization pass, to achieve a 30% relative improvement over the original diarization error rate (DER) after only two iterations of rediarization[19].

IV. CONCLUSION

In this chapter, we have presentedsome of the recent methods of speaker Diarizationfor telephone and meeting speech that have been proposed by various authors. In meeting diartization, authors used large number of new approach like bidirectional source segment based on GLR/BIC method, nonparametric clustering, the use of free energy as objective function in speaker diarization and many more. Some of them have considered new approach for finding out the speakers from overlapped speech. In telephone diarization, authors used Mealy- HMMs, Phone Adaptive Training (PAT) etc. The main aim of all the works is to improve the diarization error rate.

V. REFERENCES

[1]. Demir, C. and M. U. Dogan, "Speech-Music Segmentation System for Speech Recognition", Signal Processing and Communications Applications, 2009. SIU 2009. IEEE 17th , pp. 608-611, 2009.

[2]. Kenny, P., D. A. Reynolds and F. Castaldo, "Diarization of Telephone Conversations Using Factor Analysis", IEEE Journal of Selected Topics in Signal Processing, Vol. 4, pp. 1059–1070, 2010.

[3]. Ganchev, T., N. Fakotakis and G. Kokkinakis, "Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task", Proceedings of 10th International Conference on Speech and Computer, Vol. 2, pp. 191–194, 2005.

[4]. Hermasnsky, H., "Perceptual Linear Predictive Analysis of Speech", Journal of the Acoustical Society of America, Vol. 87, pp. 1738–1752, 1990.

[5]. Bimbot, F., J. F. Bonastre, C. Fredouille, G. Gravier and I. Magrin-Chagnolleau, et al., "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal Processing, Vol. 4, pp. 430–451, 2004.

[6]. Elie El-Khoury, Christine Sénac and JulienPinquier " Improved Speaker Diarization System For Meetings" EEE international Conference on Acoustic, speech and signal Processing , 978-1-4244-2354-5/09/$25.00 ©2009 IEEE,

[7]. DeepuVijayasenan ,"An Information Theoretic Approach to Speaker Diarization of Meeting Data" IEEE Transactions on Audio, Speech, and Language Processing, VOL. 17, NO. 7, SEPTEMBER 2009, pp. 1382-1393.

[8]. Fabio Valente, PetrMotlicek and DeepuVijayasenan ,"Variational Bayesian Speaker Diarization Of Meeting Recordings" EEE international Conference on Acoustic, speech and signal Processing, 978-1-4244-4296-6/10/$25.00 ©2010 IEEE, pp. 4954-4957.

[9]. Wei Li Yanxiong Li and Qianhua He, "Estimating Key Speaker in Meeting Speech Based on Multiple Features Optimization", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 8, No. 4 (2015), pp. 31-40.

[10]. Hayley Hung , Yan Huang, Gerald Friedland and Daniel Gatica-Perez, "Estimating Dominance in Multi-Party Meetings Using Speaker Diarization", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 4, MAY 2011, pp. 847-860.

[11]. Authors Qiao Li, Qing Fan, Yunpeng Xiao, and Weiping Ye " A Comparable Study on PNCC in Speaker Diarization for Meetings", 2010 First ACIS International Symposium on Cryptography, and Network Security, Data Mining and Knowledge Discovery, E-Commerce and Its Applications, and Embedded Systems, 978-0-7695-4332-1/10 $26.00 © 2010 IEEE, pp.157-160.

[12]. SreeHarshaYella and HervéBourlard "Overlapping Speech Detection Using Long-Term Conversational Features for Speaker Diarization in Meeting Room Conversations", IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 12, DECEMBER 2014, pp. 1688-1700.

[13]. Giovanni Soldi, Christophe Beaugeant and Nicholas Evans in their paper "Adaptive And Online Speaker Diarization For Meeting Data" , 2015 23rd European Signal Processing Conference (EUSIPCO), 978-0-9928626-3-3/15/$31.00 ©2015 IEEE, pp. 2112-2116.

[14]. SreeHarshaYella and HerveBourlard, "Information Bottleneck Based Speaker Diarization of Meetings Using Non-Speech as side Information" Acoustic Beamforming for Speaker Diarization of Meetings" , 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 978-1-4799-2893-4/14/$31.00 ©2014 IEEE, pp. 96-100.

[15]. Xavier Anguera, Chuck Wooters and Javier Hernando , "Acoustic Beamforming for Speaker Diarization of Meetings", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 15, NO. 7, SEPTEMBER 2007.pp. 2011-2022.

[16]. ItshakLapidot, Jean-Francois Bonastre and SamyBengio " Telephone Conversation Speaker Diarization Using Mealy-HMMs", Odyssey 2014: The Speaker and Language Recognition Workshop, 16-19 June 2014, Joensuu, Finland, pp. 173-178.

[17]. Simon Bozonnet, RavichanderVipperlaand Nicholas Evans "Phone Adaptive Training for Speaker Diarization", EURECOM.

[18]. RongZheng, Ce Zhang, Shanshan Zhang and Bo Xu, "Variational Bayes Based I-Vector For Speaker Diarization Of Telephone Conversations", 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), 978-1-4799-2893-4/14/$31.00 ©2014 IEEE, pp.91-95.

[19]. HoumanGhaemmaghami, David Dean and SridhaSridharan "A Speaker Rediarization Scheme for Improving Diarization in Large Two-Speaker Telephone Datasets", Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia, pp. 1272-1276.