Overview on Current Approaches in Speaker Diarization

Feng Wang

1. Introduction

In the modern age, alongside speech recognition, there is a growing need in terms of determining who is speaking during video/phone calls, recorded lectures, online business meetings, etc. Speech diarization is the field that aims to solve this issue [1]. It aims to segment the inputted speech while eliminating non-speech fragments; extract speaker identification factors from speech segments; identify each segments' corresponding speaker identities; and group segmentation according to correlated speakers [1].

Each speaker has their own unique set of characteristics in speaking, such as accent, rhythm, intonation [2]. As a result, speech signals have become a popular biometric technique [2]. There are several features, which are extracted from audio utterance utilized for identifying individual speakers, such as Mel Frequency Cepstral Coefficients (MFCC), Speaker factor, i-vector, Perceptual Linear Predictive (PLP), Neural Predictive Coding (NPC), and so on [3,4].

While these features provide potential future direction in the field. Other characteristics in a daily speech presented challenges for speech diarization. One of the key challenges in the field is speech overlapping [5,6,7,8]. In daily conversations and meetings, ranging from 8% to 17% of the speech are overlapped [6]. These overlapping speech could be deleted during segmentation or misidentified, leading to significant diarization error and lowers the accuracy of the system by 17% [5]. Another challenge is background non-speech segments. Non-speech sounds; including music, loud background chattering, animal sound; need to be removed during diarization and can introduce errors in the system. Further, for large population, it is unrealistic to train the system with all speakers' voices for identification as a close-set. The algorithm needs to be able to identify speakers who are not in the training set/system library (open-set).

2. Main Approaches

a. Signaling processing

Signaling processing has been applied in speech diarization for more than two decades. Conventional speech diarization is simplified into two stages. The first stage detects changes in the acoustic spectrum and marks those changes. The second stage identifies speakers for each acoustic spectrums [9]. One of the earliest approaches uses symmetric Kullback-Liebler (KL2) which is calculated over the 12th order PLP coefficient for the acoustic similarity between audio segments with lengths of 0.5 sec, 1.0 sec, and 4 sec [9]. This creates feature vectors that are then used for identifying non-speech and speech, and speaker clustering using the generalized likelihood ratio (BIC) [9]. This method is accurate and uses a relatively small amount of computational resources compared with other methods [9]. However, it deals with a simple broadcast scenario which is quite limiting when applied to real-life meetings [9]. In addition, since the audio segment length is preselected, if two persons talk consecutively to each other in one speech segment, this method would fail. Further, it does not take account of overlapping speech.

The probabilistic framework of expectation propagation (EP) is developed for finding accurate stopping point for conversations [10]. Individual speakers have unique identities modeled using Mel-frequency cepstral coefficient (MFCC) [10], which is a bassline acoustic feature set for speech in the 'Mel' scale [4] (1).

$$f_{mel} = 2595log_{10}(1 + \frac{f}{700}) \quad (1)$$

These unique identities allow EP to choose appropriated joint densities for margin calculation and achieve speaker diarization [10]. Another alternative approach is an adapted Gaussian mixture model, a universal background model (GMM-UBM) system on the frame to frame basis for likelihood ratio computing [11]. The system first extracts input speech features and computes against the hypothesized speaker along with speaker-independent alternative to calculate the likelihood ratio [11]. GMM-UBM then segments the audio through either internal segmentation or external segmentation [11]. While improved upon KL2, both systems still have several constraints. The most significant constrain is that for both systems each speaker is given unique identities in close-set. As a result, both approaches present a challenge for its application on a large population, when extracting individual acoustic features is too cumbersome. Further, the issue of overlapping speech is also not addressed in this approach. It appears that machine learning needs to be adopted to achieve open-set and overlapping speech diarization.

b. Machine Learning technique

I-vector is an audio feature that is widely used in speaker verification [12]. In detail, i-vector is the representation of a speech utterance in total variability space which is a low-dimensional and channel-dependent space [12]. Methods, such as cosine scoring, Variational Bayesian GMMs, mean shift, are explored to utilize i-vector for speaker diarization [13]. However, these approaches are costly and cumbersome [13]. D-vector is explored as a more efficient and accurate replacement for i-vector. D-vector is an average of speaker features which is extracted when deep neural networks are trained to group speakers [14]. D-vector is widely applied in speaker verification but is often dependent on key words detection. By adding a scoring function, the d-vector can directly perform diarization. An alternative approach is named LSTM which achieves speaker diarization by using a text-independent d-vector in combination with non-parametric spectral clustering algorism [3]. However, the clustering algorism in majority of these approaches constrains the algorism to close set training data [1]. This restrains the algorism in open-set classification. To solve this restriction, the LSTM system is further improved to unbounded interleaved-state recurrent neural network (UIS-RNN) [1] through improved clustering algorism. In UIS-RNN, RNN is an online generative process, which learns speaker information automatically [1]. This approach achieved a diarization error of 7.6% in testing [1] and is introduced by google AI as one of the most accurate open-source voice diarization system [15].

An issue that is not specified in the approaches above is overlapping speech, which leads to significant diarization errors [5]. To address this issue, a system named overlapping automatic speaker identification (OSID) is investigated [5]. Unlike previous studies, including the Gaussian Mixture Model (GMM) based system [8], multi-class Vector Taylor series (MC-VTC) system, OSID can identify more than two simultaneous speakers in the overlapping audio [5]. In the two-stage OSID (T-OSID) approach, the system first identifies the number of speakers in the speech segments, then matches the speech segments with all possible combinations with the specific number of speakers presented [5]. Direct speaker identification is achieved through a single neural network classifier in one stage OSID (S-OSID) [5]. An alternative approach is serialized output training with attention-based encoder-decoder (SOT-AED) [7]. Even though both systems can achieve significant success in identifying overlapping speakers [5,7]. It has several pitfalls for future improvements. Mainly, the system is less accurate with a high overlapping energy ratio (OER) and is solved in close-set classification [5,7].

3. CURRENT STATE OF ART OPEN SOURCE SYSTEM

The current state of art open-source application includes but is not limited to LIUM_Spk Diarization, DiarTk, and ALIZE-LIA_RAL. LIUM_Spk Diarization is a C++ toolkit that can identify open-set speakers and perform speech diarization [16]. It uses two-phase speaker segmentation utilizing generalized likelihood ratio (GLR) for speaker identification and uses Bayesian information criterion (BIC) for segmentation [16]. RiarTk is developed software using C++ and can process several feature streams simultaneously [16]. After the features; including MFCC, Modulation Spectrum (MS), Frequency Domain Linear Prediction features (FDLP); non-parametric clustering, and realignment are performed [16]. ALIZE-LIA_RAL an open-source diarization system using a non-speech GMM detection for audio classification [17]. Speech diarization is carried out using evaluative HMMs (e-HMM) [16]. Among the three systems, LIUM_Spk Diarization has the best performance when analyzing broadcast news, and TV shows [16]. However, none of the systems was able to achieve high accuracy in diarization in a more complex TV scenario, with lowest diarization error rate of 29.33% among all.

In 2018, an unbounded interleaved-state recurrent neural network (UIS-RNN) was introduced through google AI blog and is the most accurate speech diarization system in open source with a 7.6% diarization error rate [15]. By using the UIS-RNN approach, the system can identify new users and associate them with corresponding speech segments [15]. While traditionally, the open set system has lower accuracy, UIS-RNN can utilize the speaker labels and propagate the labeled data through time [15]. This is the major advantage of the system since it can identify new speakers without compensated accuracy rating. This allows the system to have major applications in patient-doctor conversations, meeting captions, broadcast documentation, etc [15]. However, this program is primarily focusing on accurate segmentation of speeches. It is not clear whether the system can identify speaker identities if overlapping speeches are presented in the audio and whether all subjects presented in overlapping speech could be identified. Further, the extent of identifying non-speech is not specified in the documentation. Since there are several types of non-speech such as music, back group noise, it is uncertain if the system can maintain high accuracy in a complicated setting.

4. FUTURE PATH

A robust speech diarization system has a wide range of applications. First, it could be used to index audio, namely broadcast, telephone call, online meetings [5,9]. With COVID-19, telephone meeting has become the new convention in business, school, and social gathering. However, telephone meetings could be harder than in-person meetings in terms of controlling progress and achieving objectives [17]. An interactive system that tracks the progress of meeting objectives and which person is talking could help participants in the meeting to engage better and avoid misunderstandings [17]. Second, it can be utilized to identify suspects in law enforcement [5,10,8]. By using voice as a biometric identity, it can also be used in authentication and surveillance [5,10]. Last but not least, the speech diarization system can be incorporated with speech recognition to take advantage of linguistic and audio cues [18]. This is particularly useful for patient and doctor appointments, especially therapy sessions and phone diagnosis appointments. In a patient-doctor scenario, each person in the meeting has a specific role with distinctive audio and linguistic cue. By utilizing linguistic and acoustic cues, accurate diarization can be achieved [18,19]. Further, by monitoring voice characteristics such as pitch and loudness, allows the system to highlight important information for participants. For doctors, this could be a certain doctor's order which they want the patient to follow strictly. For patients, this could be a trace of nervousness indicting hiding information from the doctor. Due to COVID 19, almost all doctor appointments start from a phone call examination. This system could allow better patient and doctor communication.

The last future direction particularly interests me. Through a recurrent neural network transducer (RNN-T), automatic speech recognition and speaker diarization can be incorporated together [19].

Traditionally, automatic speech recognition and speaker diarization are run parallel in the system [18]. Speech diarization breaks the segments using voice characteristics, and the deep learning model identifies the words in the speech [18]. The outputs from the two system are then combined to produce a speech transcript with speaker labels [18]. However, this approach does not utilize the linguistic and acoustic cues in conversation [18]. By incorporating automatic speech recognition and speaker diarization, the system can take advantage of both linguistic and acoustic cues, which decreased word-level diarization error by almost 10 fold to 2.2% [19]. An RNN-T model is adopted in this approach with three networks incorporated, comprised of a transcription network that maps acoustic frame; a prediction network that uses the output from a joint network to predict preceding target labels; a joint network that combined the transcription and prediction network [18,19]. The prediction label allows the system to incorporate linguistic and audio cues [19]. While this approach has no recorded open-source algorism, several existing open-source algorisms, such as ALIZE, can achieve both automatic speech recognition and speaker diarization. Further research and testing are still needed in this topic, but I believe it has huge potential in future patient-doctor appointments.

5. GitHub disclaimer

There is no GitHub code associated with this paper, due to the complicated nature of speech diarization. In depth study on both signaling processing and machine learning are required for recreation of existing open source system. Further, training system with close-set data required substantial hours to achieve desirable results.

6. Reference

[1] Zhang, Aonan, Wang, Quan, Zhu, Zhenyao, Paisley, John, and Wang, Chong. "Fully Supervised Speaker Diarization." (2019): 6301-305.

[2] Sharma, Varun, Bansal, P K, A Review On Speaker Recognition Approaches And Challenges, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 5, May – 2013

[3] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno, “Speaker diarization with lstm,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5239–5243.

[4] Hossan, M A, Memon, S, and Gregory, M A. "A Novel Approach for MFCC Feature Extraction." (2010): 1-5.

[5] Tran, Van-Thuan, and Tsai, Wei-Ho. "Speaker Identification in Multi-Talker Overlapping Speech Using Neural Networks." IEEE Access 8 (2020): 1. Web.

[6] E. Shriberg, A. Stolcke, and D. Baron, ‘‘Observations on overlap: Findings and implications for automatic processing of multi-party conversation,’’ in Proc. EUROSPEECH, Aalborg, Denmark, 2001, pp. 1359–1362.

[7] Kanda, Naoyuki, Gaur, Yashesh, Wang, Xiaofei, Meng, Zhong, Chen, Zhuo, Zhou, Tianyan, and Yoshioka, Takuya. "Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers." (2020). Web.

[8] Tsai, W. H., & Liao, S. J. (2010). Speaker Identification in Overlapping Speech. Journal of Information Science & Engineering, 26(5).

[9] Meinedo, H, and Neto, J. "Audio Segmentation, Classification and Clustering in a Broadcast News Task." 2 (2003): II-5. Web.

[10] Walsh, John MacLaren, Kim, Youngmoo E, and Doll, Travis M. "Joint Iterative Multi-Speaker Identification and Source Separation Using Expectation Propagation." (2007): 283-86. Web.

[11] Dunn, Robert B, Reynolds, Douglas A, and Quatieri, Thomas F. "Approaches to Speaker Detection and Tracking in Conversational Speech." Digital Signal Processing 10.1-3 (2000): 93-112. Web.

[12] Dehak, Najim, Kenny, Patrick J, Dehak, Réda, Dumouchel, Pierre, and Ouellet, Pierre. "Front-End Factor Analysis for Speaker Verification." IEEE Transactions on Audio, Speech, and Language Processing 19.4 (2011): 788-98.

[13] Garcia-Romero, Daniel, Snyder, David, Sell, Gregory, Povey, Daniel, and McCree, Alan. "Speaker Diarization Using Deep Neural Network Embeddings." (2017): 4930-934.Web.

[14] Variani, Ehsan, Xin Lei, McDermott, Erik, Lopez Moreno, Ignacio, and Gonzalez-Dominguez, Javier. "Deep Neural Networks for Small Footprint Text-dependent Speaker Verification." (2014): 4052-056. Web.

[15] Wang, C. (2018, November 12). Accurate Online Speaker Diarization with Supervised Learning. Retrieved September 10, 2020, from https://ai.googleblog.com/2018/11/accurate-online-speaker-diarization.html

[16] Kiktova, Eva, and Juhar, Jozef. "Comparison of Diarization Tools for Building Speaker Database." Advances in Electrical and Electronic Engineering 13.4 (2015): 314-19. Web.

[17] J. Watada and Hanayuki, "Speech Recognition in a Multi-speaker Environment by Using Hidden Markov Model and Mel-frequency Approach," 2016 Third International Conference on Computing Measurement Control and Sensor Network (CMCSN), Matsue, 2016, pp. 80-83.

[18] Shafey, L., El. (2019, August 16). Joint Speech Recognition and Speaker Diarization via Sequence Transduction. Retrieved September 11, 2020, from https://ai.googleblog.com/2019/08/joint-speech-recognition-and-speaker.html.

[19] Shafey, Laurent El, Soltau, Hagen, and Shafran, Izhak. "Joint Speech Recognition and Speaker Diarization via Sequence Transduction." (2019). Web.