DOI: 10.7763/IPEDR. 2013. V63. 1

Dialogue Transcription using Gaussian Mixture Model in Speaker Diarization

Benilda Eleonor V. Commendador+, Darwin Joseph L. Dela Cruz, Nathaniel C. Mercado, Ria A. Sagum, Diana C. Santiago, and Sharlaine Grace C. Tagnines

Polytechnic University of the Philippines, Sta. Mesa, Manila, Philippines

Abstract. Finding speaker turns and identifying the speakers is known as speaker diarization. In this study, the researchers integrate speaker diarization process with a speech-to-text task where training and/or test data may consist of two speakers using Gaussian Mixture Models. The study aims to measure how accurate the developed software in terms of missed rates, false alarm rates, speaker error rates and overall diarization error. Also, it aims to measure the accuracy of the developed dialogue transcriber software in terms of converting speech-to-text with or without proper nouns. After experimentation, results show that the Overall Diarization Error without proper nouns is 38.01% and with proper nouns got 38.16%. The dialogue transcriber is concluded to be 39.22% accurate without proper nouns while 29.95% accurate proper nouns. Based on the results, the researchers suggested enhancing the accuracy of speech-to-text system and expanding the study to more than two speakers.

Keywords: Information Retrieval, Signal Processing, Speaker Diarization, Dialogue Transcriber.

1. Introduction

The task of efficient and effective indexing and searching of growing volumes of recorded spoken documents such as broadcasts, voicemails, meetings and others requires human language technologies that cannot only transcribe speech, but also extract different kinds of non-linguistic information called metadata. Metadata includes speaker turns, channel changes and others [1].

Speech is the ordinary way for most people to communicate. Also, it can convey much information such as emotion, attitude and speaker individuality [2]. In that way, speech is the known most natural, convenient and useful means of communication. It is usable for identification because it is a product of the speaker's individual anatomy and linguistic background. In more specific, the speech signal produced by a given individual anatomy is affected by both the organic characteristics of the speaker and learned differences due to ethnic and social factors [3].

Finding speaker turns and identifying the speakers is known as speaker diarization. It is the answer to the question "who spoke when?" [4]. Its main task is to segment an audio signal into speaker-homogeneous regions without any prior knowledge of the speakers, number of speakers, text language or amount of speech present in the recording [5]. This definition simply implies that the system make use of available data in the audio recording before making decisions on speakers. Most current speaker diarization systems perform several sub-tasks which includes speech detection, speaker change detection, gender classification and speaker clustering. Broadcast News, meeting and telephone conversations are the main domains that speaker diarization is applied to [6]. It is a useful preprocessing step for an automatic speech transcription system. By separating out speech and non-speech segments, the recognizer only needs to process audio segments containing speech, thus reducing the computation time [7].

A variety of algorithms and techniques were used by researchers to increase the accuracy and performance of the diarization process. The proposed system shares a common architecture to the existing ones except it uses different algorithms for each module and as an addition, it has a speech-to-text module.

1.1. Speaker Diarization

+ *E-mail:* bennycomendador@yahoo.com

One of the applications which processes speech and converts it into useful form is Speaker Diarization. Speaker Diarization also known as the "who spoke when" task aims to group together speech segments produced by the same speaker within an audio stream [8].

Diarization is typically carried out as a three step process. The first step consists in segmenting the document into speech segments which hopefully contain speech from a single speaker – with the exception of segments containing overlapping speech. The second and third steps consist in determining the actual number of speakers and in grouping together segments from the same speaker. The Bayesian information criterion (BIC) is probably the most popular one [9].

1.2. Speech-to-Text

A speech-to-text system can be categorized by its use like command and control, dialog system, text dictation, audio transcription and other from the user's point of view. All speech recognition systems rely on at least two models: an acoustic model and a language model. All of these models can be specialized for a given language, dialect, application domain, type of speech, and communication channel to get the best transcription quality.

1.3. Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a probabilistic model for representing the presence of sub-populations within an overall population, without requiring that an observed data-set should identify the sub-population to which an individual observation belongs.

Gaussian mixture modelling is probably the most commonly used technique in audio segmentation, indexing and content analysis, in particular because of the ability of GMM to approach any distribution. Gaussian mixture models are widely used for audio classification purposes such as audio event detection, speaker recognition, gender recognition, etc. In conjunction with hidden Markov models, Gaussian mixtures can be used to simultaneously segment and classify the input audio stream.

GMM were employed due to its successful application in the speaker identification area. GMM will be used to compute segment-based likelihood given the incoming speech data. The speaker whose model gives the largest likelihood will then be identified as the target one. The Gaussian mixture speaker model maintains high identification performance with increasing population size. In the study conducted by [8], it has been concluded that the GMM provide a robust speaker representation for the difficult task of speaker identification using corrupted, unconstrained speech. The models are computationally inexpensive and easily implemented on real-time platform. Also, its probabilistic framework allows direct integration with speech recognition systems and incorporation of newly developed speech robustness techniques.

2. Dialogue Transcriber

Audio data in wav file format is fed into the activity detection module. It outputs speech segment start and end points. As soon as the segment start and end is decided, features are extracted from each speech segment. Then, it is fed into the novelty detection module. In this module, it is decided whether the segments belong to an old speaker or not. If it is a new speaker, its gender is determined in the gender identification module. After the gender of the speaker is identified, new GMM model is generated. This GMM is given a new speaker name. The words spoken by the speaker is converted to text at the same time. There are two cases for old speaker, the case if the speaker is still the previous or the speaker is in the speaker GMM. If the speaker is still the previous speaker, the system will continue to convert the speech to text. On the other hand if, the speaker is one of the previous speaker known in the system, the system will determine the speaker with the highest likelihood. The winning speaker will be then given a speaker name and convert it to text at the same time. Text format of the conversation with speech tags is produced as an output.

The system was developed in C# Programming Language, Matlab and utilizes Windows Speech Recognition for voice-to-text. The system architecture is shown in Fig. 1 and is further elaborated in the next section.

2.1. Voice Activity Detection(VAD) Module

This module aims to find the regions of speech in the audio stream and discards the non-speech (pauses) parts.

In order to analyze a voice activity, it is decomposed into two parts: the decision rule and the noise statistic estimation algorithm. These are optimized separately by applying a statistical model. A robust decision rule is derived from the generalized likelihood ratio test by assuming that the noise statistics are known a priori. For the noise statistic estimation part, a robust noise spectrum adaptation method was developed by using the soft decision information of the proposed decision rule. After marking the presence of speech in the audio, it was segmented based on its start and end points.

2.2. Feature Extraction module

In this module, features were extracted from each segment using Mel-frequency cepstral coefficient (MFCC). The MFCC process was subdivided into five phases or blocks.

- Frame blocking section- the speech waveform is more or less divided into frames of approximately 30 milliseconds.
- The windowing block- minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero.
- Fast Fourier Transform (FFT) block – converts each frame from the time domain to the frequency domain.
- Mel frequency wrapping block- the signal is plotted against the Mel-spectrum to mimic human hearing.
- Mel-spectrum – the signal is converted back to the time domain

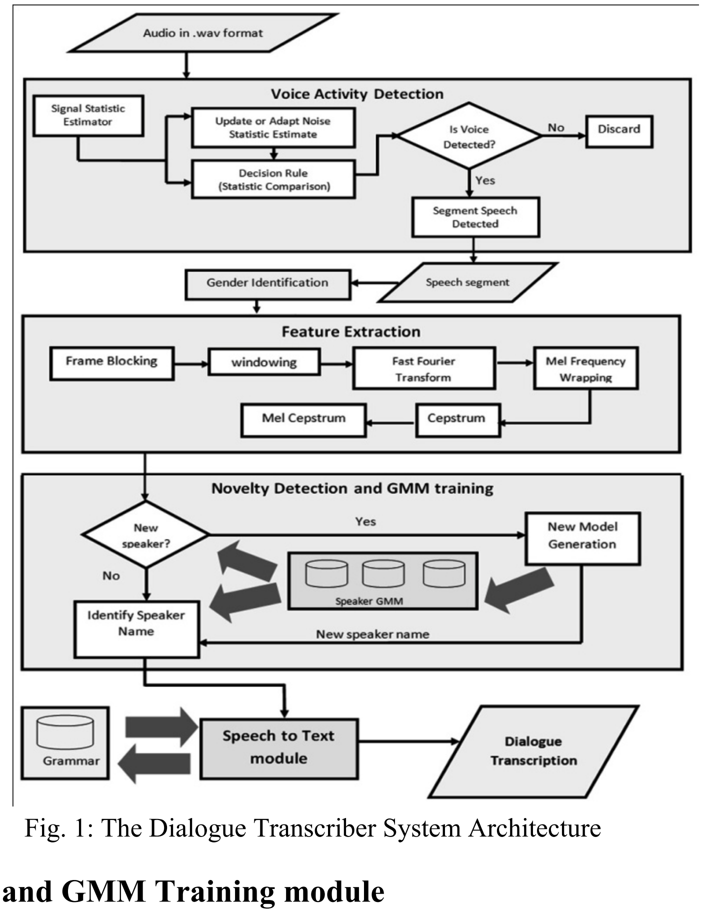

Fig. 1: The Dialogue Transcriber System Architecture

2.3. Novelty Detection and GMM Training module

The purpose of this step is to decide whether the identified speech in an audio segment comes from one of the registered system's speaker or from a new one. Maximum likelihood estimation algorithm was used to estimate the parameters our statistical models.

There are two cases for old speaker, the case if the speaker is still the previous or the speaker is in the speaker GMM set. If the speaker is still the previous speaker, the system will continue to convert the speech to text. On the other hand if, the speaker is one of the previous speakers known in the system, the system will determine the speaker with the highest likelihood. In case if there is a new speaker, the gender is identified in the gender identification module.

2.4. New Model Generation

New model is spawned by copying the parameters of the input. This new GMM is given a speaker name (the system default speaker name is Speaker1 and Speaker2) and is inserted to the system speaker GMM set.

2.5. Gender identification module

Determines the frequency of the incoming speech signal then classifies it into male and female. A typical adult male will have a fundamental frequency from 85 to 180 Hz, and that of a typical adult female from 165 to 255 Hz. [10]

2.6. Speech-to-Text module

This part converts what the speaker is saying to textual form. The speech-to-text API utilized by the study is Windows Speech Application Programming Interface (SAPI) version 5.3.

3. Implementation

3.1. Evaluation Methodology

The main metric that was used for speaker diarization experiments is the Diarization Error Rate (DER) as described and used by National Institute of Standards and Technology (NIST) in the Rich Transcription (RT). The DER error can be decomposed into the errors coming from the different sources, which includes missed rates, false alarm rate, speaker error and overall diarization error.

To identify the accuracy of the system in terms of transcribing the dialogue, Word Error Rate (WER) was used. It is the common metric of the performance of a speech recognition or machine translation system. Lower error rate implies higher accuracy of the system [11].

3.2. Results

This paper integrates a dialogue transcriber in speaker diarization system. The data needed for the study was gathered by conducting an experiment that will evaluate the system's accuracy in identifying the speakers in a conversation and evaluate its transcription output.

Table 1 shows the result of the accuracy of the speaker diarization system in terms of missed speech, false alarm rates, speaker-error rates and overall diarization error. The Overall Diarization Error without proper nouns is 38.01% and with proper nouns got 38.16%.

Table 1: Accuracy of the Speaker Diarization System.

Speaker Diarization Measure	Missed rates	False alarm rates	Speaker-error rates	Overall Diarization Error
Without Proper Nouns	3.52%	0.10%	34.39%	38.01%
With Proper Nouns	3.33%	0.16%	34.67%	38.16%

The shortcoming of the system in terms of false alarm could be attributed to the background noise. Some background noise such as music or voices not coming from the speaker in the recording could be understood by the system as a speech coming from the real speaker. Also, the system does not have an overlap detection module. In an overlapping speech, where more than one speaker may talk at a time, the false alarm is reported if any of the speakers are not detected by the system. The missed speech could be due to soft voice fed and is mistaken as a noise by the system and some speakers mumble the words as they speak. Although the words are understandable by humans, the system was not trained to those kinds of signals. Since the speaker's voice signal was used as an input to the voice activity module and is trained online, some of the trained data tagged the wrong speaker and constitutes to the propagation of errors.

Table 2: Accuracy of Dialogue Transcription.

Elements	Without Proper Nouns	With Proper Nouns
Substitutions	84	66
Deletions	29	17
Insertions	42	32
Corrects	80	40
Word Error Rate	60.78	70.05

The output reflected that the speech-to-text system is more accurate in conversations without proper nouns as shown in Table 2.

Since the speech-to-text system used during the implementation was in Windows SAPI which is designed for American Speakers, the accuracy in transcribing proper nouns was not that high when tested to Filipino Speakers. During simulation, the speakers pronounced words differently especially the names of places and persons, which were integrated by the system to different words.

The researchers recommend utilizing different speech-to-text tool aside from those of windows which is suitable for Filipino/Asian speakers.

4. Conclusion and Future Works

Throughout the implementation and evaluation phase, possible enhancements for the system and the study were established. Based on the results, the proponents viewed that the system, if further developed, should improve the speaker diarization software for higher accuracy. Device a different algorithm to improve the speech-to-text which is suitable for Filipino/ Asian speakers and also, improve the identification of proper nouns in audio processing.

5. References

[1] Konstantin Markov, S. N. (2007), Never-Ending Learning System for Online-Speaker Diarization, *IEEE* , 699-704.

[2] Toda, T. (2003, March 24), High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion, Nara Institute of Science and Technology.

[3] Suvarna Kumar, P. R. (n.d.), Speaker Recognition using GMM. International Journal of Engineering Science and Technology, 2428-2436.

[4] Margarita Kotti, V. M. (2007), Speaker Segmentation and Clustering.

[5] Reynolds, D. (2005, June 17), Automated Speaker Recognition: Current Trends and Future Direction.

[6] Themos Stafylakis, V. K. (n.d.), a Review of Recent Advances in Speaker Diarization with Bayesian Methods, 217-240.

[7] Barras, Claude. X. Z.-L. (n.d.), Improving Speaker Diarization.

[8] Sue Tranter, Douglas Reynolds, (2006), an overview of automatic speaker diarization systems. IEEE Trans. Audio, Speech and Language Processing, 1557-1565.

[9] Mathieu Ben, M. B. (n.d.), Speaker Diarization using bottom-up clustering based on a Parameter-derived Distance between adapted GMMs.

[10] Baken, R. J. (1987), Clinical Measurement of Speech and Voice. London: Taylor and Francis Ltd.

[11] (2012, September 25), Retrieved October 21, 2012, from Wikipedia: http://en.wikipedia.org/wiki/Word_error_rate