Study and analysis of Identification of Automatic speech diarization using machine learning techniques

Sayyada Sara Banu*[1]
Department of Computer Science, Jazan University, Jazan, KSA.
Rubeena[2]
Department of Computer Science, Jazan University, Jazan, KSA.
Shabana Parveen[3]
Department of Computer Science, Jazan University, Jazan, KSA.
Dr. Ratnadeep R. Deshmukhm[4]
Department of Computer Science & I.T., Dr. Babasaheb Ambedkar Marathwada University Aurangabad, India
Mohammed. Waseem Ashfaque [5`]
Department of Computer Science& I.T., Dr. Babasaheb Ambedkar Marathwada University Aurangabad, India

ABSTRACT

Many times at different situations like meetings and phone calls, knowing the speaker's identity is useful. Speaker diarization, which segments and categorizes a voice signal to the speaker identification, can be used to do this task. The i-vectors are taken out of the speech segments and used to do speaker diarization. Using both supervised and unsupervised machine learning algorithms, a model is developed for the retrieved attributes. The fresh voice segments from the speakers can be categorized using this trained model according to the corresponding speakers. Both supervised and unsupervised speaker diarization will be carried out in this study of review paper. The analysis of voice files is extremely relevant and significant in today's world of abundant voice files. In this study, the authors use supervised and unsupervised machine learning methods to propose a speaker diarization. Data processing, extraction and classification, data segmentation, and the learning phase are the system's four major segments. SVM and Multilayered- Neural Networks, two standard supervised, & the unsupervised learning technique of k-means clustering were used to train and validate the model. The authors also give an ensemble of the characteristics, which would be determined to perform much better overall, according to their findings

Keywords:

Speech diarization
Deep neural network
Features selection
Machine learning

1. INTRODUCTION

The abundance of auditory data in our environment now offers enormous potential for information access. This data can be used to retain significant conversational moments, whether they were captured during recordings of crucial technical conferences, business meetings, news broadcasts, or even straightforward telephone conversations. Additionally, it might provide additional helpful metadata that would enrich and improve transcripts significantly. Audio diary keeping is one such application. The process of identifying and classifying the many audio sources contained within an unmarked audio sequence is known as audio diarization. On the other hand, working with audio

data is time-consuming due to its lack of search ability. It would take a lot of processing power to sift through hours of audio data to locate such speaker-specific information. The use of machine learning techniques becomes the logical solution to this problem in today's developing artificial intelligence field. The model put out in this research employs supervised and unsupervised learning strategies to successfully complete the stated task.

2. LITERATURE SURVEY

With the least amount of human intervention possible, automatic speaker recognition uses algorithms and computer programs to recognize a person's speech. The identity vectors (i-vectors) technique is often used to extract voice feature characteristics from audio recordings. The features from the speech samples are then modeled using machine learning classification algorithms, such as Support Vector Machine methods (SVM) and Gaussian Mixture Model (GMM), to produce a probability score that can be compared to the known speaker [12]. Speaker diarization is a development of speaker recognition in which numerous speakers are identified and the timing of each speaker's utterance is established. By using speaker diarization, you may determine "who talked when." Both segmentation and classification are involved in this process. The identities of speakers are recovered from audio files that contain many speakers utilizing features like Mel-frequency cepstral coefficients (MFCCs) and i-vectors. A machine learning model is then trained using the identity traits that were retrieved from voice segments. New speech segments are categorized into several speaker classes using this model. Thus, each speaker's identity, instances of time, and length of speech are identified [13]. Diarization is frequently used in news broadcasts, audio recordings of meetings, and phone calls.

A useful approach for separating speech from non-speech parts and eliminating unimportant background noise is speaker diarization [3]. Speaker diarization divides a speech signal with several speakers into segments from the same speaker, creating a speaker diary involves the following three steps:

1. From the speech sample, low-dimensional i-vector features are retrieved.
2. To offer a similarity score that distinguishes different speakers, a similarity metrics tool such as the probabilistic linear discriminant analysis (PLDA) or non-linear machine learning classification probability is used [12, 14].
3. Using machine learning classification or clustering techniques, speech snippets from the same speaker are grouped together. Recent research has put forth innovative approaches to replace the conventional two-step data training process of speaker discrimination and i-vector extraction. In the novel method, embedding’s are learned from the MFCC features taken from the speech sample using deep attention models. Then, triplet loss networks, a supervised metric learning architecture, distinguish between the speakers. This trained model has been successful when tested on the CALLHOME corpus, which contains phone calls in many languages [14].

This REU project aims to use machine learning techniques to successfully perform supervised and unsupervised speaker diarization on speech samples. A part of the speech data with speaker labels is used in the supervised approach to train a SVM model that classifies the data into different speaker classes. Then, the remaining speech data is used in evaluation of system competency. K-Means clustering, used in the unsupervised technique, divides voice data without speaker labels into K number of clusters, where each cluster corresponds to a specific speaker class.

3. PROBLEM FORMULATION

This challenge is fundamentally one of grouping and recognition. Although supervised methods are also an option, traditionally these problems are treated as unsupervised learning problems. A raw audio file including recordings of talks with several speakers serves as the main input to the system. A raw audio file of a conversation serves as the system's primary input, and its output is a timeline that indicates when each speaker is taking turns. Additionally, to the input, the system can be directly provided additional metadata, such as the number of speakers, to help it and hence save time. Additionally, a portion of the voice file is provided to the system as training data for a supervised approach.

This section of the data is joined to its matching label, which in this instance is the speaker times hand-annotated label. The test set is made up of the remaining data. Figure 1 displays the system's execution flow diagram. The system is made up of four primary functional components, namely: I Data gathering ii) Extraction of features iii) Classification iv) Training and Clustering Phase, as shown by the block diagram shown. In the sections that follow, the precise operation of these modules is covered

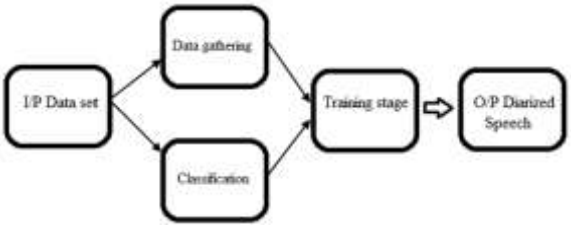

Figure 1. Work flow of the system

4. BUILDING DATASET

The original raw audio file doesn't have a lot of uses for sophisticated applications like those described in this paper. Pre-processing is required, and it is prompted in this portion.

4.1.Data gathering

Using Python's wave class and related methods, we were able to extract the audio signal from the data. Although the data was saved in stereo, we only used the monophonic signal. The 1024 window size was selected. Many approaches utilizing speech signals have determined that 1024 is the best size for speech signals. Since the only data we had was the transcript to get label information, our biggest challenge was extracting the labels for our supervised learning approaches. We initially extracted the start timings and name information from the transcript and put it in vectors. We had to match the start times with our frames and repeat the name labels the appropriate number of times until we reached the start time of a new speaker because the number of frames we would obtain if we divided our entire speech signal with the selected window is much larger than the number of name label vectors. We were able to obtain name labels for each frame of our data as a result. Additionally, we had to transform it into one hot label vector that would be suitable for employing Soft-max layer for classification in order to provide it as input labels to our neural network. The names were simply mapped to the proper vectors for implementation.

4.2. Classification

Similar to detecting change in auditory signals, speaker segmentation is also known as speaker change detection. We use a modified version of the KL distance approach to obtain the change-points by performing a single run of the full acoustic signal. Metric-based and non-metric-based speaker segmentation methods fall into two major types. The former is more well-liked and frequently applied in speaker diary algorithms. KL2 metric is a segmentation algorithm for detecting changes in speakers that is based on metrics. Through symmetrize the KL distances, the KL2 is obtained.

$$Cl2(p; q) = cl\ (p; q) + cl\ (q\ ;p)$$

4.3. Extraction of features

Machine learning systems perform better when the appropriate characteristics are extracted. This is particularly true when working with audio data, which has a variety of properties. It was crucial to choose the qualities that highlight the differences between different speakers in a dialogue. The MFCC, loudness, spectral flatness, and harmonics to noise ratio were some of the characteristics that were discovered to best meet this description. The Mel Frequency Cepstral Coefficient (MFCC) is based on the fact that human hearing cannot detect frequencies higher than 1 kHz. In other words, the essential bandwidth of the human ear is based on known fluctuation with frequency in MFCC

4.4. Volume

In this paradigm, the energy in each Bark Scale Critical Band is measured as loudness and then normalized by the total. There are 24 frequency bands that make up the bark scale critical band range, from 20 Hz to 15500 Hz. The following equation provides the conversion: Energy and frequency both affect how loud something is. Loudness presents itself as a useful attribute for diarization because the energy that a person exhibits at a particular frequency is speaker-specific for the most part.

4.5. Supporting models

Following data segmentation and feature extraction, the training phase was put into practice and is covered in this part.

4.5.1.K-means

K indicates that unsupervised learning used clustering. Developed by MacQueen in 1967, the K-means clustering technique is a popular way to automatically divide a data collection into K number of categories. It starts by choosing k initial cluster centers and refines them repeatedly as follows: I Each instance of di is paired with the nearest cluster node. 2. The mean of each cluster center Cj's constituent instances is updated. When the assignment of instances to clusters remains unchanged, the algorithm has reached its convergence. The subsequent seed assignments are made in such a way that the new seed is separated from the first by a probability weighted by the distance functions (k means++) in order to avoid the frequent mistakes of k means. This reduces the variance between classes. K's value was inputted into the system as an input for the sake of simplicity. Python's scikit-learns K-Means package was used to implement K means clustering.

4.5.2.Supervised learning

As implied by the name, supervised learning algorithms enable supervised classification in the test phase by supplying target labels. The supervised learning techniques of SVM and Multilayer Perceptron for speaker classification will be discussed in this part.

4.5.3.Deep neural network

Artificial neural networks with multiple hidden layers are known as deep neural networks. Increase the hidden-layers is primarily done towards the elimination for meticulously hand-crafted feature engineering. The requisite non-linearity for complicated logistic regression applications is also introduced. We chose to employ 3 hidden layers for our neural network because of the vast number of features we provided as input and the amount of data that was accessible. Cross entropy served as our loss function, while gradient descent was employed to train our weights. We expanded the Python code we used to create our whole model to the TensorFlow framework. Because of the intricacy of our model, using TensorFlow on a GPU allowed us to work faster. At the output of each hidden layer, we employed sof-tmax. We aim to reduce the cross entropy because labels were one of the hot vectors. To train the network, we ran the 1000 epochs. With a batch size of roughly 100, we did batch learning with a training-testing split of 0.8.

4.5.4.Support vector machine

The most frequent and well-liked approach for machine learning tasks in classification and regression is called Support -Vector Machines (SVM). It was developed by Vapnik in 1998. It functions by first mapping the input vector into a feature space with relatively higher dimensions, then finding the best separation hyper-plane there. This technique provides a series of training examples, each of which is labeled with the category to which it pertains—in this case, the speaker identification. Testing data is then categorized into the categories that were presented during training using the Support -Vector Machines algorithm. This method's strategy included the application of linear kernels. This issue is solvable mathematically. SVM is typically a binary classifier. A One Vs Rest strategy was used since the problem requires the segregation of many classes, and each test data point was either classified as belonging to or not belonging to a particular class. This was done in Python by invoking the One Vs Rest Classifier with linear svc and the scikits SVM module. The module was customized by including hyper parameters such the square-hinge loss function and L2 norm penalization.

4.5.5.Results

The outcomes of the aforementioned methods are listed in this section. 81 minutes of audio data from the Santa Barbara Corpus of Spoken English were used to train the models. The calculations were carried out using a 2.50GHz Intel Core i5-7200U CPU and an NVIDIA GeForce 940MX GPU. The outcomes are provided below. K-means The complete set of data was provided to the algorithm for clustering because it is an unsupervised learning approach. The value of K was also pre-fed into the system for convenience's sake. The clusters that had been segmented were then reorganized for temporal alignment. Following the completion of these processes, a dimensionalization error for the combined feature model of 39.14% was noted that features is ensemble and accuracy rate is 46.61.

4.5.5.1.Neural network

With the use of the abovementioned data, a neural network with two hidden layers was trained. The results are summarized as follows.

4.5.5.2.Support vector machine

The audio data was subsequently trained using the Support-Vector-Machine algorithm along the linear-kernel. To avoid overfitting, a penalization factors powered by L2 norm was incorporated.

Table 1. The result of techniques

Technique	FEATURES	ACCURACY
Neural Network	MFCC	67.1152
	Loudness	3.4022
	Spec Flat	19.1924
	Ensemble	69.2516
Support –Vector-Machines	Ensemble	91.2

5. CONCLUSION

The research problem of voice diarization and, consequently, speech recognition is open-ended. The model presented in this research is proof that machine learning techniques have been successfully applied to enhance such processes. The system has good potential even though it is not quite as reliable as modern systems. The application of something like a Recurrent Neural Network(LSTM) could help improve outcomes with the development of deep learning in this discipline in recent years. Such strategies can be used and subsequently better performance can be anticipated with increased resources like time and computer power

REFERENCES

[1] Anguera, Xavier, Chuck Wooters, and Javier Hernando. ”Acoustic beamforming for speaker diarization of meetings.” IEEE Transactions on Audio, Speech, and Language Processing 15.7 (2007): 2011-2022.

[2] Li, Chao, et al. ”Deep Speaker: an End-to-End Neural Speaker Embedding System.” arXiv preprint arXiv:1705.02304 (2017).

[3] Xavier AngueraMiro, ”Robust Speaker Diarization for Meetings”, PhD Thesis

[4] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, Speaker diarization: A review of recent research, IEEE Transactions On Audio, Speech, and Language Processing, vol. 20, pp. 356370, 2012.

[5] Jothilakshmi, S., VennilaRamalingam, and S. Palanivel. ”Speaker diarization using autoassociative neural networks.” Engineering Applications of Artificial Intelligence 22.4 (2009): 667-675.

[6] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. ”Speech recognition with deep recurrent neural networks.” Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013.

[7] Graves, Alex, and NavdeepJaitly. ”Towards End-ToEnd Speech Recognition with Recurrent Neural Networks.” ICML. Vol. 14. 2014.

[8] Vinyals, Oriol, and Gerald Friedland. ”Towards semantic analysis of conversations: A system for the live identification of speakers in meetings.” Semantic Computing, 2008 IEEE International Conference on. IEEE, 2008.

[9] Mathieu, Benoit, et al. ”YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software.” ISMIR. 2010.

[10] Shum, Stephen H., et al. ”Unsupervised methods for speaker diarization: An integrated and iterative approach.” IEEE Transactions on Audio, Speech, and Language Processing 21.10 (2013): 2015-2028.

[12] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," in *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74-99, Nov. 2015.

[13] S. H. Shum, N. Dehak, R. Dehak and J. R. Glass, "Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015-2028, Oct. 2013.

[14] C. Barras, Xuan Zhu, S. Meignier and J. L. Gauvain, "Multistage speaker diarization of broadcast news," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 15051512, Sept. 2006.

[15] H. Song, M. Willi, J. J. Thiagarajan, V. Berisha, and A. Spanias, “Triple Network with Attention for Speaker Diarization,” in*Interspeech* 2018

[16] Speaker diarization work at ICSI is collaboration between the Speech and Audio & Multimedia research groups, as well as with researchers at UC Berkeley’s ParLab and other institutions. http://multimedia.icsi.berkeley.edu/speaker-diarization/