Speaker Diarization and Identification using Machine Learning

Enhong Deng, Abhinav Dixit, Visar Berisha, Andreas Spanias
SenSIP Center and Industry Consortium
School of ECEE, Arizona State University
Tempe, AZ, USA

***Abstract*— Knowing about the identity of a speaker is helpful in scenarios such as meetings and telephone conversations. This task can be achieved using speaker diarization – segmenting and classifying a speech signal to the speaker identity. To perform speaker diarization, the i-vectors are extracted from the speech segments. A model is trained for the extracted features using supervised and unsupervised machine learning algorithms. This trained model can be used to classify the new speech segments from the speakers according to the corresponding speakers. In this paper, both supervised and unsupervised speaker diarization will be accomplished.**

Index Terms - Speaker diarization, machine learning, i-vectors

INTRODUCTION

Automatic speaker recognition uses computer programs and algorithms to identify a person's voice with minimum human involvement. Voice feature parameters are extracted from audio recordings typically using the i-vectors (identity vectors) approach. Then, machine learning classification algorithms such as Support Vector Machine algorithms (SVM) and Gaussian Mixture Model (GMM) are used to model the features from speech samples and to provide a probability score to compare against the known speaker [1]. Speaker diarization is an extension of speaker recognition where many speakers are recognized and the time of the speech from individuals is determined.

Speaker diarization answers the question "who spoke when?" This process consists of segmentation and classification. Given audio tracks that consist of several people talking, the identity of speakers is extracted using features such as Mel-frequency cepstral coefficients (MFCCs) and i-vectors. The identity features extracted from speech segments are then used to train a machine learning model. Using this model, new speech segments are classified into different speaker classes. The speaker identity, time instances and duration of speech from each speaker are therefore determined [2]. Some common places to apply diarization are broadcast news, recorded meetings, and telephone conversations.

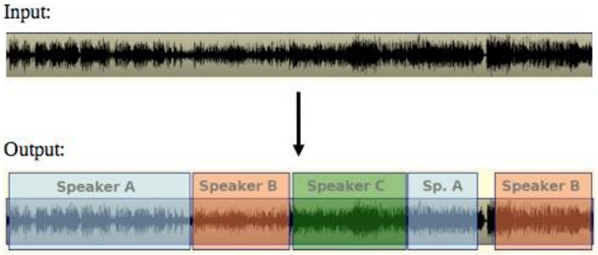

Figure 1: Identification of Speakers using Speaker Diarization [5]

Speaker diarization is a valuable tool that can separate the speech and non-speech segments and remove irrelevant background noise [3]. Fig. 1 shows how speaker diarization segments a single speech signal with multiple speakers into segments from the same speaker.

Speaker diarization comprises of following three steps:

1. Low-dimensional i-vector features are extracted from the speech sample.
2. A similarity metrics tool such as the probabilistic linear discriminant analysis (PLDA) or non-linear machine learning classification probability is used to provide a similarity score that discriminates between speakers [1, 4].
3. Speech segments that belong to the same speaker are associated together using machine learning classification or clustering algorithms.

Recent work has proposed new methods to replace the traditional two-steps data training process of i-vectors extraction and similarity metrics speaker discrimination. In the new technique, deep attention models are used to learn embeddings from the MFCC features extracted from the speech sample. Thereafter, supervised metric learning architecture called triplet loss networks differentiate between the speakers. This trained model is evaluated on a CALLHOME corpus that comprises of telephone conversations in different languages and has achieved success [4].

Our objective of this REU project is to successfully perform both supervised and unsupervised speaker diarization on speech samples using machine learning algorithms. With the supervised approach, a portion of speech data with speaker labels serves to train a SVM model that classifies the data into different speaker classes. Then, the remaining speech data is used to evaluate the performance of the trained model. With the unsupervised approach, K-Means clustering partitions speech data without speaker labels into K number of clusters, and each cluster represents a speaker class.

REFERENCES

[1] 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.

[2] 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.

[3] 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. 1505-1512, Sept. 2006.

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

[5] 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/