# Speaker Diarization and Identification Using Machine Learning


Enhong Deng, REU Student

Graduate Mentor: Abhinav Dixit, Faculty Advisors: Andreas Spanias, Visar Berisha

*SenSIP Center, School of ECEE, Arizona State University*

*Department of Speech and Hearing Science, Arizona State University*


***SenSIP Algorithms and Devices REU***

## ABSTRACT


- ❑ Speaker diarization identifies speakers in long speech recordings.
- ❑ Form speech segments and remove undesired noise and unvoiced sections.
- ❑ Form i-vectors from features extracted from speech segments.
- ❑ Train a machine learning model on extracted features.
- ❑ Classify new speech segments according to the speaker identity.


## MOTIVATION

**Speaker Recognition**

- ❑ Track the active speaker in a conversation with multiple speakers.

**Audio Indexing**

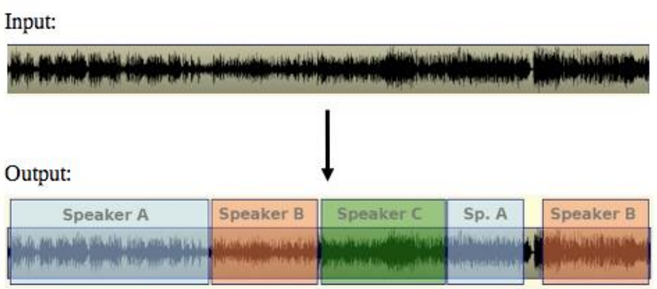
- ❑ Detect the change of speakers as a pre-processing step for automatic transcription.

**Information Retrieval**

- ❑ Examine contributions of speakers in speech recordings.

## PROBLEM STATEMENT

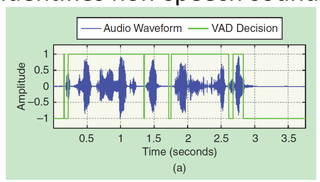
- ❑ Perform both supervised and unsupervised speaker diarization in a telephone conversation.
- ❑ Distinguish among male and female speakers to answer the question “Who speaks when?”

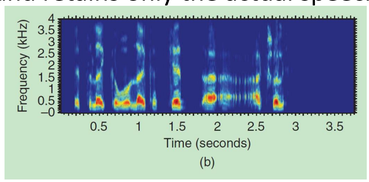



## METHODS

**Voice-Activity Detection(VAD)**

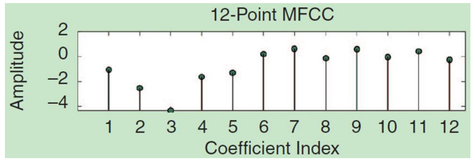
- ❑ Identifies non-speech sounds and retains only the actual speech.

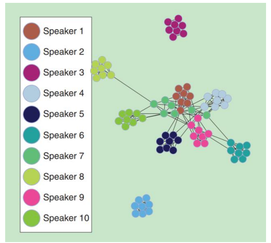



**I-Vectors**

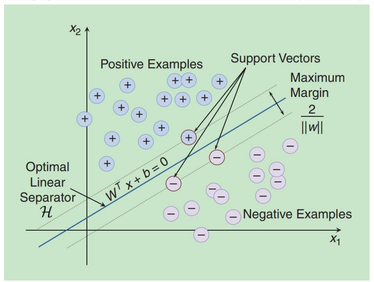
- ❑ Extracting identity information using MFCCs.
- ❑ Low-dimensional i-vectors that represent the utterances from speech.



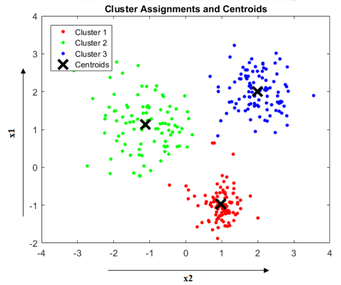

## MACHINE LEARNING ALGORITHM

**Support Vector Machines(SVM)**




- ❑ Given labeled data, SVM can be trained to develop a model capable of distinguishing among different classes.
- ❑ The trained SVM model predicts the identity of speaker in new speech data.

**K-Means Clustering**

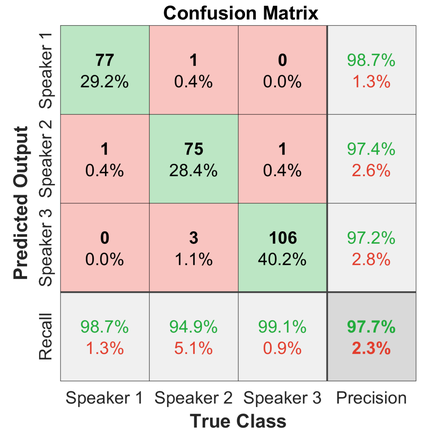



- ❑ With a known number of groups k, k number of centroids are randomly chosen.
- ❑ K-means clusters the data into k groups of clusters.

## RESULTS

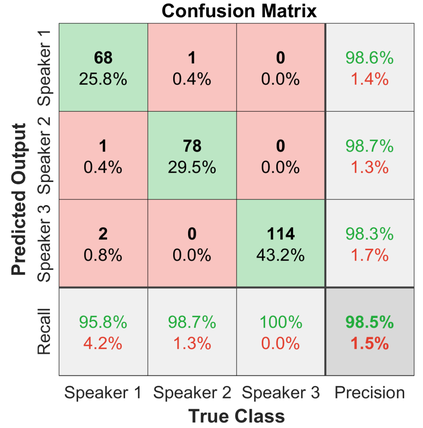
**Supervised learning**

- ❑ 97.7% accuracy in classification.
- ❑ 75% training data to generate an SVM model.
- ❑ 25% remaining data to test trained model.




**Unsupervised learning**

- ❑ 98.5% accuracy in clustering.
- ❑ All data is partitioned into three groups of clusters.
- ❑ Each cluster represents a speaker class.




- ❑ Each column of the matrix corresponds to the true class and each row corresponds to the predicted class.
- ❑ Number of correct predictions shown in green blocks, and false predictions shown in red.

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

[3] http://multimedia.icsi.berkeley.edu/speaker-diarization/


## ACKNOWLEDGEMENT


- ❑ **This project was funded in part by the National Science Foundation under Grant No. CNS 1659871 REU site: Sensors, Signal and Information Processing Devices and Algorithms.**


Ira A. Fulton Schools of Engineering
ARIZONA STATE UNIVERSITY

Sensor Signal and Information Processing Center
http://sensip.asu.edu