Using Wavelet Coefficients as Machine Learning and Deep Learning Features to Realize Speaker Diarization

Michael Chan, Kuo-Wei Lai

Brief Description

Automated speaker diarization has a lot of applications including biometric authentication and voice controlled speaker (such as Amazon Echo and Google Home). In order to identify the speaker in an audio signal, we are interested in applying what we learned in class and see how well wavelet analysis performs. Previous attempts include short time Fourier transform (STFT) and Mel-Scale Frequency Cepstral Coefficients (MFCC), but each has their drawbacks. STFT can have high computational cost and MFCC has a fixed frequency resolution. Some researchers have studied speech recognition by combining wavelet transform and artificial neural network and achieve high accuracy (98.9 %). We believe it's possible the do the same for speaker diarization. We propose a speaker diarization system that is based on wavelet analysis.

Preliminary plan and goals

We divide our plan into three main parts:

(1) *Unsupervised Learning - Identify how many different speakers in the environment*
We hope to determine the feasibility of this project by comparing the wavelet coefficients visually (e.g. plot the most important two coefficients for each training data and see their distribution, in a 2D scatter setting). First of all, we want to detect how many people speak in an audio recording. Normally in a conversation, speakers speak separately without overlapping in time (due to courtesy). Based on this assumption, we hope to segment the audio signal and label each segment with a unique speaker. We will do this by performing discrete wavelet transform (DWT) and compare wavelet coefficients.

(2) *Supervised Learning - Recognize who is speaking at what time*
In this part, we want to classify each segmentation to the corresponding speaker. DWT is our feature extraction method, and the training dataset is well labeled with the speaker identity. We will classify our audio segmentations using KNN/SVM/GMM.

(3) *Supervised Learning - Use deep learning to improve performance*
We'll use deep learning to do the same task in part(2). We expect our deep learning model to perform better than the method in part(2).

In addition to DWT, we will also implement different approaches (MFCC and wavelet packet) to do the same tasks. By doing so, we can compare the performance of these methods and understand the strengths and weakness of each method.

Reference:

[1] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, ‘‘Speaker diarization: A review of recent research,’’ IEEE Trans. Audio Speech Lang. Process., vol. 20, no. 2, pp. 356–370, Feb. 2012.

[2] Muda Lindasalwa, Begam Mumtaj and Elamvazuthi I.,“ Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”, Journal Of Computing, Volume 2, Issue 3, pp 138-143, ISSN 2151-9617, March 2010.