Novel Features for Speaker Recognition

Phyu Phyu KHING (3121063)
Supervisor : A/Prof. E.Ambikairajah
Assessor : A/Prof R.Ramer

Jiamin Karen KUA (3137442)

Abstract

This thesis implements two novel features extraction algorithms at the frontend of Automatic Speaker Recognition (ASR) system. Recent studies suggest that Frequency Modulation (FM) components are required to improve the performance of ASR in realistic environments. A range of techniques to extract AM and FM features are investigated. Among them, Comprehensive Modulation Spectra (CMS) and Frequency Amplitude Modulation Encoding (FAME) algorithms are chosen to extract AM and FM features from a given speech. Speaker identification and verification experiments are done at ASR backend to see the relative improvement of incorporating FM features into the system. The simulated results verify that AM+FM features outperform the AM only features by an average of 9%.

1. System Overview

Generally, an Automatic Speaker Recognition system consists of two main modules: features extraction and features matching. Features extraction is the frontend process that converts speech waveform to a parametric representation called feature vectors. Feature matching is the backend process that identifies the speaker by comparing extracted features from the unknown speaker’s voice with a set of enrolled speakers. Based on system requirement threshold, accept or reject decision is made.

2. Novel Features Extraction Algorithms

2.1 Comprehensive Modulation Spectra (CMS) Algorithm

CMS is a new feature extraction algorithm working on temporal evolution of spectrum, which manifested in slowly varying modulation properties of the speech signal. There are two pathways in CMS module to extract AMS and FMS separately. AMS is defined as log energy variation of the speech signal, whereas FMS is defined as intensity weight average instantaneous frequency change with time.

2.2 Frequency Amplitude Modulation Encoding (FAME)

The FAME algorithm was developed to decompose a signal into slowly varying amplitude and frequency modulation.
- The signal is divided into sub-bands and the frequency modulation is extracted from each.
- FM extraction in each sub-band is performed by a quadrature oscillator shown above.
- The means of the sub-band FM signals for each frame are taken as the modulation features.

3. The Back End: Gaussian Mixture Model (GMM)

In this system, GMMs are used to produce a probabilistic model for the AM and FM feature vectors extracted from each algorithm. GMM is a parametric representation of a probability density function, based on weighted sum of multivariate Gaussian distributions. In all algorithms, GMMs are used to train near-ideal condition speech from the TIMIT (Texas Instruments Massachusetts Institute of Technology) database. A testing procedure is then used to determine the speaker verified/identified percentages with a given utterance. The final system uses 32 GMMs for the training of the AM and FM features from CMS, FAME and cepstral coefficients from MFCC.

4. Final Results

5. Conclusions

- Novel features based on amplitude and frequency modulation have been proposed.
- Different features and number of speakers are compared.
- The simulated results in Table 1 shows that AM+FM features does outperform the AM only features for both proposed algorithms by an average of 9%.
- ASR performance improves if the combination of Amplitude Modulation and Frequency Modulation features is done in a systematic way.