Dynamic Feature Extraction for Speech Signal Based On MUSIC and Modulation Spectrum

Han Zhiyan and Wang Jian

College of Engineering, Bohai University, No. 19, Sci-tech Road, High & New Technological and Industrial Development District, Jinzhou, China, 121000

(Received 19 August 2015; accepted 30 June 2016)

A novel dynamic feature extraction algorithm is proposed to help improve speech recognition robustness in noise environmental conditions. Owing to the modulation spectrum having time-frequency agglomeration performance, according to different reflections in the modulation spectrum for interference and speech signals, we first calculate the Multiple Signal Classification (MUSIC) spectrum and then get the modulation spectrum. We then filter the modulation spectrum signal. For the filter signal, use a 32 frames signal is used as a processing unit to get the modulation spectrum energy vector. This reveals the close correlation between the speech signal frames and can well reflect speech dynamic characteristics. Finally, the cepstrum coefficients are extracted as the feature parameter. It not only adequately reflects speech dynamic characteristics, but also has lower sensitivity for the speech environment. The effectiveness of the feature is discussed in view of the class separability and speaker variability properties. We evaluated the feature under different kinds of noise (white noise, pink noise, street noise, and panzer noise) and different signal-to-noise ratios (-5 dB, 0 dB, 5 dB, 10 dB, and 15 dB). The experimental results show that the novel feature has good robustness and computational efficiency under low signal-to-noise ratios and plays a very good foreshadowing role in latter speech research.

NOMENCLATURE

MFCC:	Mel Frequency Cepstral Coefficient
MUSIC:	Modulation Spectrum; Multiple Signal Classification
LPC:	Linear Prediction Coefficients
LSP:	Linear Spectrum Pair
FFT:	Fast Fourier Transform
LDA:	Linear Discriminant Analysis
DM:	Determinant Measure
HMM:	Hidden Markov Model
BPNN:	Back-Propagation Neural Networks
SNR:	Signal-to-Noise Ratio

1. INTRODUCTION

Research on the robustness of speech recognition was still a challenging task, especially in the development of core speech processing algorithms. One example was feature extraction from speech signals. Namely, extracting features that could reflect speech characteristics from the speech waveforms. It not only could reduce the number of calculations and storage, but could also filter out useless and redundant information and was one of the most fundamental and important aspects of speech recognition. Some time domain features, such as amplitude feature, short-time frame average energy, short-time frame zero crossing rate, short-time autocorrelation coefficient, et cetera appeared. With the development of recognition technology, we found that the stability and separating capacity for the time domain features were not good, so we began using the frequency domain feature as a speech feature, such as pitch period, formant frequency, linear prediction coefficients (LPC), linear spectrum pair (LSP), cepstrum coefficient, and so on. Among them, the MFCC, which was based on the auditory model, was a widely-applied feature at present. However, once these features were used in the noise environment, their performance droped sharply.[1–4]

The features mentioned above reflect the static feature of speech signal, while a dynamic feature was part of speech diversity, which was different from a stationary random process. It had time correlation, and revealed the close relationship between speech signal (pre and post). We could get the dynamic feature by differential parameters and acceleration parameters for the static feature. However, differential parameters and acceleration parameters could not fully dig out dynamic information. Therefore, studying the dynamic feature of the speech signal was an inevitable trend to improve the performance of speech recognition.

Estimating the time-varying spectrum was a key first step in the speech feature extraction.[5–10] MFCC was computed by applying a Mel-scaled filter bank either to the short-time Fast Fourier Transform (FFT) magnitude spectrum or to the short-term LPC-based spectrum. However, both FFT and LPC-based spectrum were very sensitive to noise contamination. Eigenvector-based methods, such as MUSIC, were popular in sinusoidal frequency estimation due to their high resolution and less prior information. Moreover, MUSIC algorithm had well noise restraining ability. We adopted the MUSIC spectrum instead of the traditional method.