HOKKAIDO UNIVERSITY

Title	A Study on Robust Speech Recognition with Time Varying Speech Features [an abstract of dissertation and a summary of dissertation review]
Author(s)	Mufungulwa, George
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第12944号
Issue Date	2017-12-25
Doc URL	https://hdl.handle.net/2115/68113
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	George_Mufungulwa_abstract.pdf, 論文内容の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称　博士（情報科学）　氏名　George Mufungulwa

学 位 論 文 題 名

A Study on Robust Speech Recognition with Time Varying Speech Features
（音声の時変特徴量を用いた雑音ロバスト音声認識に関する研究）

Speech feature extraction algorithms have become popular. Speech features can be used for various applications: biometric recognition, speech recognition, speaker identification, and so on. In these applications, a good speech feature can be obtained using Mel frequency cepstrum Coefficients (MFCC), Linear Predictive Coding (LPC), Time varying LPC (TVLPC), Perceptual Linear Predictive (PLP) among others. This thesis focuses on TVLPC among feature extraction algorithms to improve the robustness of automatic speech recognition (ASR) systems against various multiplicative and additive noises. Time varying speech features (TVSF) are implemented in ASR with the aim of improving the recognition accuracy with a number of small set of reference speech databases. The significance of the study is based on the fact that both additive and multiplicative noises cause great performance degradation of ASR systems, thereby limiting the speech recognition accuracy in real environments. For this reason, feature correction, compensation and normalization approaches are considered in order to improve the robustness of a speech recognition system.

The performance degradation is partly due to statistical mismatch between trained acoustic model of clean speech features and noisy testing speech features. For the purpose of reducing the feature-model mismatch, corrective, compensation as well as normalization techniques are employed during training and testing of speech features. In order to achieve improved system performance, normalization in modulation spectrum domain is used to remove non-speech components over a certain frequency range using an enhanced running spectrum analysis (RSA) as a band pass filter. In comparison to other noise reduction techniques used in this study for speech recognition, the advantage of the enhanced RSA filter is its adaptable parameters, that is, the first and second pass band frequencies can easily be adjusted accordingly. In addition, speech feature enhancement using dynamic range adjustment (DRA) aiming at correcting the difference between clean and noisy speech features by normalizing amplitude of speech features is utilized. For the purpose of channel normalization, cepstrum mean subtraction (CMS) is used in this study.

In order to find some statistically relevant information from incoming data, it is important to have mechanisms for reducing the information of data segment in the audio signal into a relatively small number of parameters, or features. In this study, there are two subsystems for feature extraction.

Two alternative time varying speech features (TVSF) methods are proposed. The first alternative is directly converted time varying linear prediction (TVLPC) based MFCCs, while the second alternative

mel filtering and logarithmic transformations are applied to short time windowed time varying coefficients before converting to cepstrum coefficients in place of direct-converted TVLPC speech features. In the final analysis, the mel filtering with logarithmic transformations TVLPC based MFCCs was utilized.

Thirteen models were initially formulated, models 0 to 12 of feature vectors. In this study, models are formulated as follows: First, model 0 makes use of fast Fourier transform (FFT-based MFCC) only and is referred to as a conventional method. Models 1 to 12 are formulated as follows: FFT based MFCC coefficients consisting of 38-dimensional feature vectors are concatenated with TVLPC based MFCC coefficients. The number of TVLPC-based MFCC coefficients appended to FFT-based MFCC coefficients represents the effective model number. For example, appending a single TVLPC-based MFCC coefficient to 38 coefficients of FFT based MFCC result in model 1 and appending 2 TVLPC based MFCC coefficients to 38 coefficients of FFT-based MFCC result in model 2 and so on. In final analysis models 0 to 4 were utilized. In the simulations the performance of conventional approach and proposed approach are evaluated on clean speech as well as noisy speech in MATLAB (R2014a) software.

The following speech data sets are utilized: similar pronunciation phrases, set (a): /denki/, /genki/ and /tenki/ and set (b): /kyuu/, /juu/ and /chuu/, set (c): 13 phrases /ohayou/,/konnichiwa/,/kombanwa/, /onamaewa/,/genki/, /tanoshiine/,/arigato/,/denki/,/tenki/,/kyuu/,/juu/,/migi/, and /hidari/, uttered by elderly persons and set (d): 142 Japanese common speech phrases that include all other speech data of set (a), set (b) and set (c) for male speakers. Initially, a database of 40 male speakers is made available for the study. Prior to the commencement of experiments, the initial database is split into two parts, the first part consisting of 30 speakers and is used for the front-end feature extraction and HMM training. The second part consisting of 10 speakers is utilized in the testing stage. The speech sample is 11.025 KHz and 16-bit quantization. For standard speech information processing, the frame concept has been applied. The 256 (23.2ms) sample point length frame is first defined and using this frame, a short time speech waveform is extracted. For the short time speech waveform, a speech power spectrum is calculated as a typical speech analysis. The frame is shifted with 128 (11.6ms) points and then many short time speech waveforms can be obtained.

This thesis is divided in seven chapters beginning with the introduction that describes the background of digital processing of speech, study motivation, thesis overview and contributions. Chapter 2 introduces the standard method for speech recognition and the steps in conventional FFT-based. Time varying speech feature (TVSF) algorithm of direct TVLPC and TVSF algorithm of mel based TVLPC are discussed. Chapter 3 discusses the voice activity detection (VAD) method with short term energy, short term autocorrelation and zero-crossing rate (ZCR). Chapter 4 discusses the influence of additive and multiplicative noises. Modulation spectra and noise circumstances, the running spectrum filter and running spectrum analysis algorithm are discussed. In addition, the limitation of band-pass filtering

is highlighted. Chapter 5 evaluates the performance of proposed time varying speech features with HMM. Experiments are conducted for the two sets of 3 similar phrases, 13 phrases uttered by elderly people and the 142 common Japanese speech phrases. Chapter 6 compares the proposed time varying speech features to those of the conventional approach. Chapter 7 summaries the above research and gives a conclusion to highlight the research significance. Finally, some possible works for future research are briefly described.

The numerical performance of speech recognition on similar pronunciation phrases, phrases uttered by elderly persons and common Japanese phrases are evaluated. For data set (a): /genki/,/denki/ and /tenki/ the recognition accuracy results indicate that models 1 and 3 show an average improvement of 6.45% and 3.33% to the conventional method respectively. For data set (b): /kyu/, /juu/ and /chuu/ results show that model 1 is slightly better at an average of 0.90%. For data set (c): 13 phrases uttered by elderly persons model 1 performs better at an average of 0.36% recognition accuracy compared to the conventional method.

Enhanced RSA has been applied for the following frequency components; RSA Type (a) : 1 Hz to 7 Hz, RSA Type(b): 1 Hz to 15 Hz, RSA Type(c): 1 Hz to 30 Hz, RSA Type(d): 1 Hz to 35 Hz and RSA Type(e): 1 Hz to 40 Hz and each band-pass is evaluated on 15 types of noises at 5 dB, 10 dB, 15 dB, 20 dB and 25 dB SNR. RSA Type (d) showed better performance than the rest. Among the five noise levels, at 54.04%, 5 dB was the best.

It is found that concatenating FFT-based MFCC with the TVLPC-based MFCC to create time varying speech features (TVSF) can improve the speech recognition capability for some models with HMM implementations. The proposed method may be a simpler solution for speech recognition applications that requires further improvements.