CONTENT

CONTENT ..... **i**

LIST OF FIGURES ..... **iv**

LIST OF TABLES ..... **vi**

ABSTRACT ..... **vii**

CHAPTER I **1**

INTRODUCTION **1**

1.1 History of ASR Technology 2
1.2 Characteristic of Speech Recognition Systems 4
1.3 Motivation of Research 5
1.4 Problem Formulation 7
1.5 Position of Research 8
1.6 Research Purpose 14
1.7 Benefits of Research 15
1.8 Contribution of Research 16

CHAPTER II **17**
LITERATURE REVIEW AND BACKGROUND THEORY **17**

2.1 Literature Reviews 17
2.1.1 Acoustic Feature Extraction 17
2.1.2 Pattern Classification in ASR 19
2.2 Acoustic Feature Extraction Techniques in ASR 20
2.2.1 Linear Predictive Coding (LPC) 21
2.2.2 Linear Preciction Cepstrum Coefficient (LPCC) 22
2.2.3 Perceptual Linear Prediction (PLP) 23
2.2.4 Mel Frequency Cepstral Coefficient (MFCC) 24
2.2.5 Relative Spectral Filtering (RASTA) 26
2.3 Noise Robustness Techniques in MFCC 29
2.3.1 Discrete Wavelet Thresholding Denoising 30
2.3.2 Double Density Discrete Wavelet Transform (DD-DWT) 32
2.3.3 Noise Robust RASTA-MFCC 34

2.3.4 Perceptual Linear Prediction Cepstral Coefficients (PLPCC) 35
2.4 Speech Recognition Methodologies 36
2.4.1 Template-Based Approaches 36
2.4.2 Knowledge-Based Approaches 37
2.4.3 Dynamic Time Wrapping (DTW) based Approaches.....37
2.4.4 Support Vector Machine (SVM).....38
2.4.5 Statistical-Based Approaches.....39
2.4.5.1 Hidden Markov Model.....39
2.4.5.2 Gaussian Mixture Model – Hidden Markov Model.....44
2.4.5.3 Deep Learning Deep Neural Network – DNN-HMM 46
2.4.5.4 Deep Learning – Recurrent Neural Network 48
2.5 Creating Decoding Graph 50
2.6 Kaldi Speech Recognition Toolkit.....51
2.7 Performance Metrics 52
CHAPTER III 54
RESEARCH METHODOLOGY 54
3.1 Equipment and Materials 54
3.2 Roadmap of Research 59
3.2.1 The First Stage 59
3.2.2 The Second Stage.....60
3.2.3 The Third Stage.....61
3.3 Typical Speech Recognition Architecture 61
3.4 Psychoacoustic Model 65
3.4.1 Auditory Masking 67
3.4.1.1 Simultaneous Masking.....67
3.4.1.2 Non-simultaneous Masking 68
3.4.2 Implementation of Simultaneous Masking 69
3.4.2.1 Finding High-Resolution Spectral Estimate 70
3.4.2.2 Finding Tonal and Non-tonal Components 70
3.4.2.3 Determination of Valid Tonal and Non-tonal Maskers ..71
3.4.2.4 Determination of Individual Masking Thresholds.....72

3.4.2.5 Figuring the Global Masking Threshold 73
3.4.2.6 Figuring the Minimum Masking Threshold 74
3.5 Structure of Perceptual Masking Effect-based Gammatone Frequency Cepstral Coefficients (PGFCCs) Feature Extraction 75
3.5.1 Frame Blocking and Windowing 76
3.5.2 Modeling Simultaneous Masking Effect 78
3.5.3 Filterbank Analysis 78
3.5.3.1 Gammatone Filterbank 79
3.5.3.2 Bark-scale Filterbank 81
3.5.4 Discrete Cosine Transform 82
3.5.5 Short-term Energy 83
3.5.6 Cepstral Mean Variance Normalization (CMVN) 84
3.6 Acoustic Feature Recognition Engines 84
3.6.1 GMM-HMM Modeling 85
3.6.2 Frame Level Cross-entropy DNN-HMM Modeling 87
3.6.3 Sequence Discriminative Training of DNN 89
3.7 Evaluate the Method 92
CHAPTER IV 100
IMPLEMENTATION AND EXPERIMENTAL RESULTS 100
4.1 Results and Discussion 100
4.2 Statistical Analysis for Evaluated Performance using ANOVA 111
4.3 Comparison Target of Modified MFCC 116
4.3.1 Discrete Wavelet Denoising into Conventional MFCC 117
4.3.2 Double Density Discrete Wavelet Transform into MFCC 118
4.3.3 Relative Spectral Filtering (RASTA) into MFCC 119
4.3.4 Robustness of Perceptual Linear Prediction 120
4.4 Evaluation on Open Noise Environment 123
4.5 Summary 125
CHAPTER V 127
CONCLUSION AND FUTURE WORK 127
REFERENCES 129

LIST OF FIGURES

Figure 1.1 Speech system milestones over the past 45 years 3
Figure 1.2 Fishbone diagram of inaccurate text transcription in ASR 7
Figure 2.1 The detail process flow of LPC feature extraction technique 22
Figure 2.2 Process flow of LPCC feature extraction technique 22
Figure 2.3 Detail steps of PLP feature extraction technique 23
Figure 2.4 The detail process of MFCC feature extraction technique 24
Figure 2.5 Structure of triangular Mel filterbank 25
Figure 2.6 Process flow of RASTA technique 27
Figure 2.7 Process of discrete wavelet transform denoising scheme 31
Figure 2.8 A 3 Channel perfect reconstruction filterbank in DD-DWT..... 32
Figure 2.9 Process flow of MFCC-RASTA feature extraction technique 35
Figure 2.10 The block diagram of PLP technique..... 35
Figure 2.11 Compare time series data using Dynamic Time Warping 37
Figure 2.12 Classification of two linear separable classes 38
Figure 2.13 Trellis of Hidden Markov Model (HMM) 39
Figure 2.14 Architecture of Gaussian Mixture Model based HMM 44
Figure 2.15 Architecture of hybrid Deep Neural Network-HMM 46
Figure 2.16 An architecture of DNN..... 47
Figure 2.17 Architecture of Recurrent Neural Network..... 49
Figure 2.18 Simplified view of different component in Kaldi toolkit..... 51
Figure 3.1 Graphical representation of a utterance under subway noise..... 57
Figure 3.2 Graphical representation of a utterance under babble noise 57
Figure 3.3 Graphical representation of a utterance under car noise 58
Figure 3.4 Graphical representation of a utternce under exhibition noise 58
Figure 3.5 Roadmap of proposed research..... 59
Figure 3.6 Detail procedures of the first stage 60
Figure 3.7 Detail procedures of the second stage..... 60
Figure 3.8 Detail procedures of the third stage 61
Figure 3.9 Overview of typical speech recognition system 62
Figure 3.10 Hearing range and human sound perception..... 66

Figure 3.11 Typical phenomena of auditory masking 67
Figure 3.12 Paradigm of simultaneous masking 68
Figure 3.13 Paradigm of non-simultaneous or temporal masking 69
Figure 3.14 Tonal and non-tonal masker on a linear frequency scale 72
Figure 3.15 Individual masking threshold on a linear frequency scale 73
Figure 3.16 Global Masking Threshold 74
Figure 3.17 Minimum Masking Threshold for each subband 75
Figure 3.18 Overview of proposed perceptual masking effect-based gammatone frequency cepstral coefficients (PGFCC) feature extraction technique 76
Figure 3.19 Blocking multiple short-time frames of a speech signal 77
Figure 3.20 Hamming window of 64 sample points 77
Figure 3.21 64 Channel frequency response of Gammatone filterbank 81
Figure 3.22 Bark-scale filterbank with 64 Trapezoidal filters 82
Figure 3.23 Architecture of hybrid DNN-HMM system 88
Figure 3.24 Process flow diagram of proposed robust ASR system 93
Figure 3.25 Block diagram of conventional MFCC feature extraction 94
Figure 3.26 Process flow diagram of proposed PGFCC 95
Figure 3.27 Process flow of feature recognition process using GMM-HMM and DNN-HMM acoustic model training scheme 97
Figure 4.1 Mean word error rate (%) for validation of choosing number of hidden layers in cross-entropy DNN-HMM acoustic model building 103
Figure 4.2 Overall validation on PBFCC and PGFCC at different SNRs 106
Figure 4.3 Results of overall noise validation on proposed PBFCC and PGFCC under different types of noise situation 106
Figure 4.4 The detail accuracy (%) of MFCC and PGFCC under different noise situations and SNRs using sequence discriminative DNN-HMM model 109
Figure 4.5 Overall validation of MFCC and proposed PGFCC using sequence discriminative DNN-HMM training 110
Figure 4.6 Recognition accuracy (%) for comparison targets of converntional MFCC using sequence discriminative DNN-HMM training 122
Figure 4.7 Comparison of mean accuracy (%) on open noise situations 125

LIST OF TABLES

Table 1.1 Summary of previous researches in modified MFCC approaches 12
Table 2.1 Comparative study of LPC, LPCC, MFCC, PLP and RASTA..... 28
Table 2.2 The set of asymmetric analysis filters and synthesis filters..... 33
Table 3.1 Summary of Aurora-2 connected digits speech corpus 56
Table 4.1 Recognition performance (%) of conventional MFCC on different noise and SNRs using GMM-HMM and cross entropy DNN-HMM training..... 102
Table 4.2 Recognition accuracy (%) of proposed PGFCC on different noise and SNRs using GMM-HMM and cross entropy DNN-HMM training..... 104
Table 4.3 Recognition accuracy (%) of PBFCC on different noise and SNRs using GMM-HMM and cross entropy DNN-HMM training..... 105
Table 4.4 Recognition accuracy (%) of MFCC using state-level minimum Bayes risk criterion (sMBR) sequence discriminative acoustic model 107
Table 4.5 Recognition accuracy (%) of proposed PGFCC using sMBR criterion sequence discriminative acoustic model..... 108
Table 4.6 Outlines of Two-Way ANOVA table 112
Table 4.7 Statistically significant evaluation on the performance of three acoustic models GMM-HMM, cross-entropy and sequence discriminative DNN-HMM 113
Table 4.8 Statistically significant evaluation on performance of feature extraction with conventional MFCC and proposed PGFCC..... 115
Table 4.9 Evaluation of MFCC and PGFCC in terms of execution time 116
Table 4.10 Accuracy of DWT using sequence discriminative DNN-HMM..... 118
Table 4.11 Accuracy of DD-DWT using sequence discriminative DNN-HMM119
Table 4.12 Accuracy of RASTA using sequence discriminative DNN-HMM.. 120
Table 4.13 Accuracy of PLPCC using sequence discriminative DNN-HMM .. 121
Table 4.14 Accuracy (%) proposed PGFCC feature extraction on open noise testing set using sequence discriminative DNN-HMM training..... 124