Psychoacoustic model for robust speech recognition

Luo, Xue Wen

2008

Luo, X. W. (2008). Psychoacoustic model for robust speech recognition. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/41749>

<https://doi.org/10.32657/10356/41749>

Psychoacoustic Model for Robust Speech Recognition

XUEWEN LUO

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Master of Engineering

2008

Acknowledgements

I am greatly indebted to my supervisor, Associate Prof. Soon Ing Yann for accepting me as his student and kindly providing guidance and advice throughout the whole project. His comments have been of greatest help at all times.

Gratitude also goes to Associate Prof. Yeo Chai Kiat and Prof. Koh Soo Ngee for sharing their valuable knowledge, experience and skills generously.

Finally, I also extend my sincere thanks to my family members for their support and understanding.

Table of Contents

Acknowledgements..... i

Table of Contents ii

Summary vi

List of Figures viii

List of Tables x

List of Tables x

Chapter One

Introduction..... 1

1.1 Motivation..... 1

1.2 Objectives 4

1.3 Major Contributions of This Thesis..... 5

1.4 Organization of Thesis..... 5

Chapter Two

Overview of Automatic Speech Recognition 7

2.1 Development of Automatic Speech Recognition..... 7

2.2 Fundamentals of Speech Recognition..... 9

2.3 Hidden Markov Model..... 12

2.4 Feature Extraction..... 15

2.4.1 Pre-emphasis..... 17

2.4.2 Windowing..... 19

2.4.3 Fast Fourier Transform 19

2.4.4 Mel Filter Bank Analysis..... 20

2.4.5 Discrete Cosine Transform	22
2.4.6 Energy Measure	22
2.4.7 Derivative Calculation	23
Chapter Three	
Indonesian Database and Recognizer	24
3.1 Database Preparation	24
3.1.1 Introduction to Indonesian	24
3.1.2 Recording and Transcription.....	29
3.1.3 Dictionary Preparation	31
3.2 Overview of Current Indonesian Database	33
3.3 Indonesian Recognition System.....	37
3.3.1 Training.....	38
3.3.2 Testing.....	41
3.4 Tests and Results.....	41
3.4.1 DR Dependent Testing in Phoneme Level.....	41
3.4.2 DR Independent Testing in Phoneme Level	42
3.4.3 Overall Testing in Phoneme Level	43
3.5 Subjective Analysis over the Recognition Outputs	43
3.5.1 Speed of the Speech	44
3.5.2 Multiple Dictionary Entries	44
3.5.3 Informal Indonesian Pronunciations	45
Chapter Four	
Simultaneous Masking Modeling for ASR.....	47

4.1 Introduction to Psychoacoustic	47
4.2 Introduction to Simultaneous Masking.....	48
4.3 Simultaneous Masking Modeling	51
4.4 Test and Results	54
4.4.1 Overview of the Testing Database.....	54
4.4.2 Front End for Baseline System	55
4.4.3 Recognizer Back End.....	55
4.4.4 Lateral Inhibition without Weighting	55
4.4.5 Lateral Inhibition with Weighting	56
4.4.6 Symmetrical and Asymmetrical Lateral Inhibition.....	57
4.4.7 Lateral Inhibition with Temporal Integration	57
Chapter Five	
Temporal Masking Modeling for ASR.....	59
5.1 Introduction to Temporal Masking	59
5.2 Forward Masking Modeling	63
5.2.1 Type I Forward Masking Modeling.....	63
5.2.2 Type II Forward Masking Modeling.....	67
5.3 Test and Results	67
5.3.1 Results of Type I Modeling	68
5.3.2 Results of Type II Modeling.....	68
Chapter Six	
Integrated and 2D Psychoacoustic Modeling for ASR	69
6.1 Integrated Psychoacoustic Modeling	69

6.1.1 Proposed Feature Extraction Front End 69

6.1.2 Cepstral Mean Normalization (CMN) 70

6.2 Test and Results 71

6.3 2D Psychoacoustic Modeling 77

6.3.1 Introduction to 2D Psychoacoustic Modeling..... 77

6.3.2 2D Psychoacoustic Modeling and Parameters 78

6.4 Test and Results 82

Chapter Seven

Conclusion and Recommendations 84

7.1 Conclusion 84

7.2 Recommendations for Further Research..... 85

7.2.1 Integration of Backward Masking 85

7.2.2 Improvement of the 2D Psychoacoustic Modeling..... 85

7.2.3 Further Testing on more Databases 86

7.2.4 Application of the Auditory Model to Indonesian Recognizer..... 86

Author’s Publications..... 87

Bibliography 88

Summary

This thesis presents a detailed study on psychoacoustic modeling for feature extraction for robust speech recognition. In an automatic speech recognition (ASR) system, feature extraction is critical to determining the recognizer's performance. The most popular feature vectors for ASR are Mel Frequency Cepstral Coefficients (MFCC). However, it is also well known that its performance drops dramatically under noisy condition. One of the objectives of this thesis is to improve the robustness of a recognizer.

Compared to an ASR system, human is good at tolerating background noise, hence psychoacoustic modeling of human hearing system is investigated and integrated into speech features extraction process of a speech recognizer to increase the robustness of it. The psychoacoustic phenomena that have been studied include both simultaneous masking in the frequency domain and temporal masking in the time domain.

Firstly, simultaneous masking effect (lateral inhibition) and temporal masking effect (forward masking) have been modeled and integrated with a speech recognizer independently. Moreover, another masking effect called temporal integration is introduced as an auxiliary to lateral inhibition. The valuation tests are carried out on AURORA 2, which is a widely used English digits database. The results show that either of simultaneous and temporal masking effects contributes to recognition rate increase separately.

Secondly, an integration version of both simultaneous masking effect (lateral inhibition) and temporal masking effect (forward masking) is proposed and tested. A recognizer with the integration version of these masking effects is compared with that with cepstral mean normalization (CMN), which is a commonly used feature enhancement algorithm for ASR. The results show that the proposed auditory modeling algorithm registers a higher improvement in recognition rate as compared to the CMN.

Last but not least, an innovative psychoacoustic model which is a 2D filter is proposed. Be different from other conventional psychoacoustic modeling, which cover the masking effects no other than in frequency and time domain (two directions only), the proposed 2D psychoacoustic model covers masking effects in a 2D surface. This pioneering algorithm shows competitive recognition rate improvement over the AURORA 2.

Besides, an Indonesian telephony conversational database with 7 hours data is reported in this thesis. The recording and transcribing of audio files and dictionary preparation are described in thesis. A preliminary phoneme level recognition system for continuous Indonesian language has been set up.

List of Figures

Figure 1 Key Milestones in Speech Recognition over the Past 50 Years [5] 8

Figure 2 Speech Recognition Process..... 10

Figure 3 HTK Software Architecture [14]..... 12

Figure 4 Process of Converting a Speech Waveform to Feature Vectors..... 16

Figure 5 MFCC and PLP Parameterization 17

Figure 6 Frequency Response of Pre-emphasis Filter when $\alpha=0.97$ 18

Figure 7 Windowing of MFCC..... 19

Figure 8 Mel Scale Filter Bank..... 21

Figure 9 Speech Recognition System Implemented by HTK 37

Figure 10 a Topology Model for Phoneme..... 38

Figure 11 Sketch of Lateral Inhibition..... 51

Figure 12 Simplified Lateral Inhibition Model..... 52

Figure 13 Average Forward Masking Data (circles) and Standard Deviation (error bars), together with the Model Line Fit (lines) as a Function of Masker Level across Five Octaves, with Probe Delay of 15, 30, 60 and 120 ms as Parameter [49]..... 61

Figure 14 Example of Recovery Process for Forward Masking [49] 62

Figure 15 Comparison between Traditional MFCC and Proposed MFCC with Masking Effects 70

Figure 16 Recognition Rate (%) under Car Noise 74

Figure 17 Recognition Rate (%) under Train Station Noise..... 75

Figure 18 Superiority of MFCC+LI+TI+FM under Clean Training Condition 76

Figure 19 Superiority of MFCC+LI+TI+FM under Multi Training Condition..... 77

Figure 20 Comparison of MFCC, Integrated MFCC+LI+TI+FM and 2D Psychoacoustic
Models..... 79

Figure 21 1D Mexican-hat Filter 80

List of Tables

Table 1 Indonesian Phoneme Set	27
Table 2 Rules of Transcription Format	30
Table 3 Dialect Regions of Indonesia	33
Table 4 Indonesian Speech Corpus Data	34
Table 5 Phoneme Level Recognition Results for Each DR	42
Table 6 Phoneme Level DR Independent Recognition Results	43
Table 7 Overall Results in Phoneme Level	43
Table 8 Sample RECOUNT File for Speed of the Speech	44
Table 9 Sample REOUT File for Multiple Dictionary Entries	45
Table 10 Sample REOUNT File for Informal Indonesian Pronunciations	45
Table 11 Set A Recognition Rate (%) of Lateral Inhibition without Weighting	56
Table 12 Set A Recognition Rate (%) of Lateral Inhibition with Weighting	56
Table 13 Set A Recognition Rate (%) of Symmetrical and Asymmetrical Lateral Inhibition	57
Table 14 Set A Recognition Rate (%) of Lateral Inhibition with Temporal Integration ..	58
Table 15 Model Parameters for Forward Masking	64
Table 16 Centre Frequencies for Mel Scaled Filter Banks	64
Table 17 Model Parameters for Each Mel Filter Bank	66
Table 18 Set A Recognition Rate (%) of Type I Forward Masking Modeling	68
Table 19 Overall Recognition Rate (%) of Type II Forward Masking Modeling	68
Table 20 Set A Recognition Rate (%) of Type II Forward Masking Modeling	68
Table 21 Overall Recognition Rate (%) for AURORA 2	72

Table 22 Recognition Rate (%) of MFCC+LI+TI+FM VS. MFCC (39) 72

Table 23 Recognition Rate (%) of MFCC+LI+TI+FM VS. MFCC (39) 73

Table 24 2D Psychoacoustic Modeling Parameters..... 81

Table 25 Recognition Rate (%) of 2D Filter VS. MFCC (39)..... 83

Table 26 Recognition Rate (%) of 2D Filter VS. MFCC (39)..... 83

Table 27 Recognition Rate (%) of 2D Filter VS. MFCC+LI+TI+FM 83

Chapter One

Introduction

1.1 Motivation

There are many novel motivators for researchers to improve on the speech recognition system. Potential applications of speech recognition include:

- Integrating it with a computer's operating system, so that the user can work with computer hands-free and eyes-free. The added advantage is that dictation of words and paragraphs is normally much faster than typing;
- Multi-tasking can also be achieved through the use of speech recognition, an example will be the dropping of fire retardant from a helicopter, when your hands are busy with flight control;
- Bringing convenience to the physically handicapped by using voice to control various devices in daily life;
- The output of a speech recognition system can also undergo further linguistic processing in order to achieve speech understanding.

Other than the convenience motivators, the demand of the market is another major motivator for speech recognition. According to figures provided by the industry, satisfying the needs of consumers and businesses by simplifying customer interaction, increasing efficiency, and reducing operating costs can be achieved by incorporating

speech recognition in a wide range of applications. Furthermore, Allied Business Intelligence (ABI) [1] has released that the speech recognition market would increase to \$897.8 million in 2003, up from \$677 million in 2002, and over the longer term, the market is forecasted to grow to \$5.3 billion by 2008. This is due to its growing reputation, as accuracy rates will be much improved and new applications to meet users' needs are currently being developed. However, the speech recognition market didn't increase as fast as expected, one of the main reasons is that challenges of a recognizer are not easy to be overcome.

The main challenge in speech recognition is robustness according to Dr. Lawrence Rabiner's notes [2]. The challenge of robustness is introduced by taking into account different classes of recognition mode.

- The recognition mode can be either speaker dependent or speaker independent. Some recognition systems require speaker enrollment, which means a user must provide samples of his or her speech to train the acoustic models before the system can be used, which called speaker dependent system. Whereas other systems are speaker independent, which means no beforehand training is necessary.
- Even for the same speaker, the speaker may sounds different, such as after drinking or catching a cold.
- Speaking style can differ when reading from a script or carrying out a natural conversation. Natural conversation contains disfluencies, and it doesn't follow the formal grammar, so it is much more difficult to recognize than speech read from a script.

- The input mode can be either isolated words or continuous speech. An isolated word recognition system requires that the speaker pause briefly between words, whereas a continuous speech recognition system does not.
- The surrounding acoustic environment can range from very quiet, e.g. in soundproof room, to extremely noisy, e.g. public area. Generally speaking, the noisy speech is harder for recognition system, since system may not be smart enough to distinguish the background noise and the unvoiced parts. If the background noise is so high so that all intelligibilities of speech may lost at all.
- The vocabulary size can be smaller than a hundred words or larger than a few thousands words. Recognition is generally more difficult when vocabularies are large, since the recognition can be only done in sub-word level and a combination process is compulsory during the recognition output process, which may reduce the accuracy. Language models or artificial grammars can be used to restrict the combination process to avoid the reduction.

There are also some other accessorial challenges of robustness, such as the capability of the recognition system to add out-of-vocabulary words into word lexicons, and the capability of the system to generate language models for new tasks. Hence there is much room for improvement in the speech recognition area.

Other than those motivations and challenges mentioned above for a recognizer in the automatic speech recognition system, the database for evaluation is also important. A lot of English, Mandarin, Japanese and other languages' databases are available now from

the Linguistic Data Corporation. However, being the mother tongue of the 4th largest population in the world, Indonesian has very few databases currently. One text corpus that has been published was collected by Information and Language Processing Systems (ILPS), Informatics Institute, University of Amsterdam [3]. A recent paper [4] also mentioned that there are three other types of Indonesian data resources available in both text and speech forms. The first two resources were developed or processed by the R&D Division of PT Telekomunikasi Indonesia (R&D TELKOM) in collaboration with Advanced Telecommunications Research Institute International (ATR). The third one was developed by ATR under the Asian Speech Translation Advanced Research (A-STAR) project in collaboration with Badan Pengkajian Dan Penerapan Teknologi (BPPT, Language Technology Laboratory of Indonesia).

1.2 Objectives

Since human auditory system can tolerate much more background noise than an automatic speech recognition (ASR) system. This thesis presents a detailed study on psychoacoustic modeling, which includes simultaneous masking in frequency domain and the temporal masking in time domain for feature extraction for robust speech recognition. Hence the main objectives of the research are:

- To implement and evaluate simultaneous masking (lateral inhibition);
- To implement and evaluate temporal masking (forward masking);
- To implement and evaluate a integration version with both simultaneous masking and temporal masking;
- To implement and evaluate a pioneering 2D psychoacoustic model.

1.3 Major Contributions of This Thesis

The main contributions of this thesis include the following:

- Setting up an Indonesian telephony conversational database including almost 7 hours speech.
- Setting up a phoneme level Indonesian recognizer base on the current Indonesian database.
- Investigation on the psychoacoustic phenomena of human for speech recognition. Several psychoacoustic models have been proposed and implemented, to increase the robustness of a speech recognizer. One of which is the integration of both the simultaneous masking effect and forward masking effect into the speech feature vectors.
- Another contribution is pioneering the use of a 2D filter to simulate the simultaneous masking in frequency domain and temporal masking in time domain concurrently.

1.4 Organization of Thesis

This report consists of seven chapters. In Chapter 2, an overview of automatic speech recognition system is given. Each element of a recognition system will be discussed in this chapter. More emphasis will be given to the feature extraction front-end, as it is the main focus of the proposed algorithm. In Chapter 3, the process of setting up an Indonesian database and an overview of the existing database will be presented. The recognition results by an preliminary phoneme level recognizer for the Indonesian continuous language will be shown. In Chapter 4, a proposed model for simultaneous

masking and the evaluation test results will be presented. A description of a commonly used English database – AURORA 2 will also be given as this is used for benchmarking the proposed algorithm. In Chapter 5, a proposed model for temporal masking and the evaluation test results will be presented. In Chapter 6, the integration of both simultaneous masking and temporal masking will be presented. In addition, a novel 2D psychoacoustic model with both simultaneous and temporal masking effects will be proposed. The evaluation on both AURORA 2 and the Indonesian database will be performed. Conclusion and recommendation will be given in Chapter 7.

Chapter Two

Overview of Automatic Speech Recognition

A brief introduction to automatic speech recognition is given in this chapter. The development process of automatic speech recognition is presented in Section 2.1. The basic elements of an automatic speech recognition system are described in Section 2.2. Some of the basic theory and equations of the Hidden Markov Model (HMM) will be reviewed in Section 2.3. The current commonly used methods for feature extraction in speech recognition system are listed in Section 2.4.

2.1 Development of Automatic Speech Recognition

Speech recognition was brought to the attention of researchers by AT&T Bell Laboratories in the 1940s by developing a primitive device which can recognize speech. Over time, the speech recognizer has improved steadily generation by generation. In the 1960s, researchers turned their focus to the development of more complex speech recognition systems. At the first stage, the system required the discrete input speech, verbal stimuli punctuated by small pauses. However, in the 1970s, continuous speech recognition was introduced, which does not require the users to pause between words. A handful of milestones in speech recognition over past 50 year are highlighted in Figure 1 [5]. The technology of speech recognition is still under refinement for commercial usages.

Figure 1 Key Milestones in Speech Recognition over the Past 50 Years [5]

State-of-the-art speech recognition needs to be narrowed down by exploiting the constraints of the task. Moreover, different technologies are sometimes appropriate for different tasks. For example, when the vocabulary is small, the entire word can be modeled as a single acoustic unit. Such an approach is not practical for large vocabularies, where word models must be built up from sub-word units.

After working on isolated-word, speaker-dependent systems for many years, the community has moved towards very-large-vocabulary (20,000 words and above), high-perplexity, speaker-independent continuous speech recognition since 1992. Databases with vocabulary of thousands of words are intended primarily for the dictation application. The continuous speech recognition allows people to speak in a natural manner.

Even though much progress is being made such that a context-dependent recognition system may work reasonably well, machines are still a long way from recognizing conversational speech. The best context-dependent system achieved an error rate of 7.2% on read sentences drawn from North America business news in 1994 [7]. Word recognition rates on telephone conversations in the Switchboard corpus are around 50% [8]. It will be many years before unlimited vocabulary and speaker-independent continuous dictation capability is technically feasible.

Microsoft released speech recognition tools with the windows operating system – WindowsXP Service Pack 2. The users need to train the recognizer themselves before the recognition process, and the training process will take nearly half an hour. No exact number on Microsoft's website shows the accuracy of their recognition tool. Subjective surveys with ten persons on their satisfaction of using the Microsoft recognition tool had been conducted by the author. The result shows that they feel the Microsoft's recognizer works well in isolated words recognition, but they are not satisfied with the performance of its continuous speech recognition.

2.2 Fundamentals of Speech Recognition

Speech recognition is the process of converting speech, captured by a microphone or a telephone, into a set of words. The standard speech recognition process is shown in Figure 2. Firstly, the speech waveform input will be converted into feature vectors by feature analysis. Then, these features that represent the speech information will be provided to a speech recognizer. The most commonly used recognizer for large vocabulary recognition – Hidden Markov Model (HMM) is shown in Figure 2 below.

There are also some other recognizers used for speech recognition, for example, the Dynamic Time Warping (DTW) [9], a hybrid of DTW and HMM [10], the Artificial Neural Network (ANN) [11] and a hybrid of HMM and ANN [12].

Figure 2 Speech Recognition Process

Speech recognition systems can be characterized by many parameters. An isolated-word speech recognition system requires that the speaker pause briefly between words, whereas a continuous speech recognition system does not. Spontaneous or extemporaneously generated speech contains disfluencies, and is much more difficult to recognize than speech read from script. Some systems require speaker enrollment - a user must provide samples of his or her speech to train the acoustic models before the systems can be used, whereas other systems are said to be speaker-independent, which means no training is necessary. Some of the other parameters depend on the specific task. Recognition is generally more difficult when vocabularies are large, since the recognition can be only done in sub-word level and a combination process is required, which may reduce the accuracy. Language models or artificial grammars can be used to restrict the combination of words.

The performance of a speech recognition system is measurable. Perhaps the most widely used measurement is recognition rate, defined as:

$$RecognitionRate = \frac{N - D - S}{N} \times 100\% \quad (2.1)$$

Where N is the total number of labels or words, D is the number of deletion, and S is the number of substitutions in the test file.

At the last stage of the recognition system, the system will give out the recognition output. Sometimes, not all the utterances that are processed by the speech recognizer are correct, so a confidence measure is always provided as an indicator.

In this thesis, HMM based software called HTK (Hidden Markov Model Toolkit) is used for those experiments. HTK is the Hidden Markov Model Toolkit developed at the Speech Vision and Robotics Group of the Cambridge University Engineering Department (CUED) in 1989. This toolkit aims at building and manipulating Hidden Markov Models (HMMs). HTK consists of a set of library modules and tools available in C source code form, and is primarily used for speech recognition research. The tools provide sophisticated facilities for speech analysis, HMMs training, testing and results analysis [13, 14]. Figure 3 provides the software structure of the HTK and shows its input and output interfaces.

Figure 3 HTK Software Architecture [14]

2.3 Hidden Markov Model

Due to its rich mathematical structure and theoretical basis and high practicability for applications, Hidden Markov Models (HMM) are widely applied to automatically speech recognition [15].

Hidden Markov Model is an extension of Markov Chain. In a regular Markov model, each state is visible and the state transition probabilities from one state to the other are the only parameters. In a Hidden Markov Model, the state is not visible directly, however each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by a HMM provides information about the hidden sequence of states. The challenge of HMM is to determine the hidden sequence of states based on the sequence of tokens.

An HMM consist the following elements:

- The number of states in a model, N . The set of all states is denoted by $S = \{S_1, S_2, \dots, S_N\}$ and at time t , the model is at the state q_t .

- The number of observations generated by a state, M . The set of observation symbols is denoted by $V = \{v_1, v_2, \dots, v_M\}$.

- The state transition probability, $A = \{a_{ij}\}$, and

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \quad 1 \leq i, j \leq N$$

- The observation probability distribution in a state, $B = \{b_j(k)\}$, and

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j], \quad 1 \leq j \leq N \text{ and } 1 \leq k \leq M$$

- The initial state, $\pi = \{\pi_i\}$, and

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N$$

Therefore a complete specification of an HMM requires specification of two model parameters (N and M), specification of observation symbols (V), and the specification of the three probability measures (A , B and π). Usually, a HMM is represented by $\lambda = (A, B, \pi)$ for convenience.

By setting appropriate values of N , M , A , B , and π , the HMM can generate an observation sequence ($O = O_1 O_2 \dots O_T$, where each observation O_t is one of the symbols from V , and T is the total number of observations in the sequence) as follows:

- Choose an initial state $q_1 = S_i$ according to initial state distribution π .

- Set $t = 1$.
- Choose $O_t = v_k$ according to the symbol probability distribution in state S_i , i.e., $b_i(k)$.
- Transit to a new state $q_{t+1} = S_j$ according to the state transition probability distribution for state S_i , i.e., a_{ij} .
- Set $t = t + 1$, and repeat the previous two steps, until $t = T$.

As the given form of HMM above, there are three basic problems of interest for the applications:

- Problem 1: Given the observation sequence $O = O_1 O_2 \cdots O_T$, and a model $\lambda = (A, B, \pi)$, how to efficiently compute $P(O|\lambda)$, the probability of the observation sequence?
- Problem 2: Given the observation sequence $O = O_1 O_2 \cdots O_T$, and a model $\lambda = (A, B, \pi)$, how to choose a corresponding state sequence $Q = q_1 q_2 \cdots q_T$, which is the optimal in some meaningful sense?
- Problem 3: How to adjust the model parameter $\lambda = (A, B, \pi)$ to maximize $P(O|\lambda)$?

All of these three problems are involved in automatic speech recognition process. An isolated-word recognition process can be used as an example for explanation: an N -state HMM is used for each word of a vocabulary of W . Assume that the coding is done by using a spectral codebook with M unique spectral vectors. Hence each observation is the

index of the spectral vector closet (in some spectral sense) to the original speech signal. Thus, for each vocabulary word, a training sequence consisting of a number of repetitions of sequences of codebook indices of the word (by one or more talkers) is obtained. The first step is to build individual vocabulary word models. This step is done by using the solution to Problem 3 to optimally estimate model parameters for each word model. To develop an understanding of the physical meaning of the model states, the solution to Problem 2 is used to segment each of the word training sequences into states, and then study the properties of the spectral vectors that lead to the observation occurring in each state. Finally, after a set of W HMMs has been optimized (training process), recognition of an unknown word is done by the solution to Problem 1 to score each word model based on the given test observation sequence (the codebook of each test word), and select the word with highest likelihood.

2.4 Feature Extraction

In an ASR system, the speech waveform input is converted into feature vectors representing the acoustic information by feature extraction analysis. Figure 4 gives a flow diagram of the conversion process. Then, these features that represent the speech information will be provided to a speech recognizer. Feature extraction is a critical element in speech recognition since it is the first step of the recognition process and generates the parameters on which the recognition algorithm is based. At the feature representation level, robust techniques are developed to account for distortions due to noise, stress and microphone differences. If the feature parameters are inappropriate, the recognition performance is naturally limited. A good speech feature extractor can ease the burden on the speech recognizer. The environment noise could distort the speech

waveform and affect the intelligibility of speech seriously. Hence, the recognition rate under noisy condition is much worse than that under clean condition. Feature extraction in a classification problem is also about reducing the dimensionality of the input-vector while maintaining the discriminating power of the signal. In the process of analysing the features, no assumption is made on the noise characteristics, although this may be viewed as either an advantage or a disadvantage.

Figure 4 Process of Converting a Speech Waveform to Feature Vectors

The most widely used speech feature vector is based on the Mel Frequency Cepstral Coefficients (MFCC) [17]. The other commonly used feature vector is Perceptual Linear Prediction (PLP) coefficient [18 and 19] which is based on the Linear Predictive Coefficient (LPC) [20, 21, and 22]. Figure 5 shows the process of MFCC and PLP

parameterization. Since many databases use MFCC as their baseline results, discussion will be focus on the MFCC in this thesis.

Figure 5 MFCC and PLP Parameterization

2.4.1 Pre-emphasis

An important pre-processing step in the analysis of speech signals is the pre-emphasis of high frequencies. This is done because the amount of energy contained in the high energy components is small compared to low frequencies. The frequency bandwidth of a speech is about 16kHz, with most of the speech energy below 7kHz. As the speech spectrum has a drop off 6-dB/octave, the high frequency components have smaller amplitude with respect to the low frequency components. This results in a high dynamic range within the speech bandwidth. Hence, a pre-emphasis at high frequencies is required to make all the components similar in amplitude.

Pre-emphasis is usually performed by filtering the signal with a FIR filter whose transfer function in time domain is:

$$y(n) = x(n) - a \cdot x(n-1) \quad (2.2)$$

where a is close to 1, with typical values around 0.97.

This pre-emphasis filter is a first order FIR. Figure 6 shows the frequency response of this filter. Generally speaking, it suppresses the low frequency components and amplifies the high frequency components. The output signal of the pre-emphasis filter has a relatively lower dynamic range.

Figure 6 Frequency Response of Pre-emphasis Filter when $a=0.97$

2.4.2 Windowing

Speech signal is non-stationary, so it is necessary to process it frame by frame. Within each frame, the signal can be assumed to be near-stationary. The frame length is typically 20 to 40 ms. The simplest technique for framing is the rectangular window. However, a rectangular window has high side-lobes in its frequency spectrum, which is called spectral leakage. In order to solve this problem, windows which taper off at both ends are generally used. In speech feature extraction, the most commonly used window shape is the Hamming window, defined by

$$w(n) = 0.54 - 0.46 \cos\left[2\pi \frac{n-1}{L}\right] \quad 0 \leq n \leq L \quad (2.3)$$

The windowed frames are overlapped, typically by 50% as showed in Figure 7.

Figure 7 Windowing of MFCC

2.4.3 Fast Fourier Transform

MFCC uses the Discrete Fourier Transform (DFT) to convert the speech signal from the time domain to the frequency domain, which is defined as

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-j \frac{2\pi nk}{N}\right), \quad k = 0, 1, 2, \dots, N-1 \quad (2.4)$$

where N is the number of samples.

In practice, Fast Fourier Transform (FFT) is normally applied instead to reduce the computational load. In FFT, N is taken to be the nearest number to the power of 2. For example, a speech signal sampled at 16kHz, and windowed at length of 25 ms, the number of sample in each frame is 400. N is taken to be 512 by appending zeros at the end of the sample sequence for FFT calculation.

2.4.4 Mel Filter Bank Analysis

The Mel scale is based on an empirical study of the human perceived pitch or frequency. The scale is divided into the unit of “Mel”. The test subjects in the study started out hearing a frequency of 1000Hz, which was labelled 1000 Mels for reference. The testees were then asked to change the frequency until they perceived the frequency to be twice the reference. This frequency was then labelled 2000 Mels. The test was then repeated with half the frequency, 1/10, 10 times and so on, labelling these frequencies 500 Mels, 100 Mels, and 10000 Mels. Based on these results a mapping of the normal frequency scale to the Mel scale was possible.

The Mel scale is generally speaking a linear mapping below 1000Hz and logarithmically spaced above. The mapping relationship between Hz and Mel is represented by

$$Mel(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (2.5)$$

Davis and Mermelstein (D&M) coined the term ‘Mel frequency cepstral coefficients’ (MFCC) in 1980[17] when they combined non-uniformly spaced filters with the discrete cosine transform (DCT) as a front-end algorithm for automatic speech recognition (ASR). The algorithm can be summarized as: a signal passes through a triangular filter bank, spaced on a linear-log frequency axis, and the energy output from each filter is log-compressed and transformed via the DCT to cepstral coefficients.

Figure 8 Mel Scale Filter Bank

The filter bank is normally comprised of 10 linearly-spaced centres below 1kHz and 10 log-spaced filters above 1kHz. The base of each triangular filter is determined by the centre frequencies of the neighbouring filters, and all filters are of the same height. The spacing of the filter bank mimics the warped frequency characteristics of the human auditory system as described above. The Mel filter bank is composed of triangular filters on a linear logarithmic scale. The base of each triangle defined by the centre frequency of adjacent filters thus the base of each filter overlaps with its neighbours by 50%.

Davis and Mermelstein (D&M) coined the term ‘Mel frequency cepstral coefficients’ (MFCC) in 1980[17] when they combined non-uniformly spaced filters with the discrete cosine transform (DCT) as a front-end algorithm for automatic speech recognition (ASR). The algorithm can be summarized as: a signal passes through a triangular filter bank, spaced on a linear-log frequency axis, and the energy output from each filter is log-compressed and transformed via the DCT to cepstral coefficients.

Figure 8 Mel Scale Filter Bank

The filter bank is normally comprised of 10 linearly-spaced centres below 1kHz and 10 log-spaced filters above 1kHz. The base of each triangular filter is determined by the centre frequencies of the neighbouring filters, and all filters are of the same height. The spacing of the filter bank mimics the warped frequency characteristics of the human auditory system as described above. The Mel filter bank is composed of triangular filters on a linear logarithmic scale. The base of each triangle defined by the centre frequency of adjacent filters thus the base of each filter overlaps with its neighbours by 50%.

After FFT, all the frequency components are grouped together by the filter banks. In each frequency band, the frequency components are weighted according to the filter frequency response and then accumulated to a value that represents the total energy of the band. This step may result in some loss of information from the original signal, although it is widely accepted that such a step is helpful in extracting the useful components in speech signals for recognition.

2.4.5 Discrete Cosine Transform

The Discrete Cosine Transform (DCT) is used to de-correlate the filter bank's coefficients. Spectral coefficients are transformed into cepstral coefficients according to:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right) \quad \text{for } i = 1, 2, \dots, N \quad (2.6)$$

where m_j is the filter bank coefficient and N is the number of filter banks. The logarithm simply converts the multiplications into additions. The number of cepstral coefficients is usually taken to be around 12. Hence N ($N = 20$ to 24 usually) Mel filter bank coefficients are compressed into 12 cepstral coefficients during this step.

2.4.6 Energy Measure

To extend the cepstral coefficients from DCT, an energy term can be appended. There are two ways of calculating the energy coefficient: log energy (logE) and the zero coefficient (c0).

The log energy is calculated in time domain as $E = \log \sum_{n=1}^N s_n^2$, where the s_n represent speech samples and N is the total sample number within each frame. The zero coefficient is calculated by

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right) \quad \text{for } i = 1, 2, \dots, N \quad (2.7)$$

when $i = 0$. In fact, the zero coefficient is a scaled sum of the Mel filter bank coefficients. Either c_0 or $\log E$ can be appended to the cepstral vector. The two coefficients can be used at the same time as well but any further improvement will be marginal.

2.4.7 Derivative Calculation

The cepstral coefficients are normally referred as static features, that is, they only carry information from a given frame. Hence, dynamic features are introduced to enhance the frame representations. The delta features are also known as velocity features. Second derivatives such as delta-delta or acceleration features can also raise the performance of a speech recognition system [27]. After DCT and appending energy term, the 13-dimensional static feature vector (MFCC_E or MFCC_0) is appended with its first and second derivative (MFCC_E_D_A or MFCC_0_D_A). The 39-dimensional speech features are widely used during the recognition process. Furthermore, delta coefficients and acceleration coefficients are computed by using

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \quad (2.8)$$

where d_t is a delta coefficient at time t computed in terms of the corresponding static coefficients $c_{t-\Theta}$ to $c_{t+\Theta}$. Θ is set to $\Theta = 2$ in most cases.

Chapter Three

Indonesian Database and Recognizer

In an ASR system, the training process requires a large amount of speech utterances (tens of hours) and each utterance must be labeled with a phone or word level transcription. The speech recording and transcription is tedious and time consuming. Since Indonesian has the 4th largest population in the world, and little work has been done on the Indonesian speech recognition, one of the objectives of this thesis is to set up an Indonesian speech database, which currently contains 7 hours of manually transcribed data. The preparation process of the Indonesian database is given in Section 3.1. The overview of the current Indonesian database is given in Section 3.2. In order to evaluate and compare the proposed feature extraction algorithm with current techniques reported in the literature, the evaluation is carried out using AURORA 2, which is a widely used English database. An introduction to AURORA 2 is given in Section 3.3.

3.1 Database Preparation

3.1.1 Introduction to Indonesian

Bahasa Indonesia (Indonesian) is the official language of Indonesia. The language is a standardized dialect of the Malay language. However there are some variations in the pronunciation and vocabulary due to the presence of Dutch and Javanese words in Indonesian vocabulary [16]. As mother tongue of 4th largest population in the world, Indonesian has very few databases currently. One text corpus which has been published,

is collected by the Information and Language Processing Systems (ILPS), Informatics Institute, University of Amsterdam [28]. A recent paper [29] also mentioned that there are three other types of Indonesian data resources available in both text and speech forms. The first two resources were developed or processed by the R&D Division of PT Telekomunikasi Indonesia (R&D TELKOM) in collaboration with Advanced Telecommunications Research Institute International (ATR). The third one was developed by ATR under the Asian Speech Translation Advanced Research (A-STAR) project in collaboration with Badan Pengkajian Dan Penerapan Teknologi (BPPT, Language Technology Laboratory of Indonesia).

Only 7% of the population in Indonesia uses the official language as their daily language [30], with most of them living in Jakarta. On the other hand, more than 200 million people speak the Indonesian language in various ways. However, the native speakers rarely use the correct standard version of the language in daily conversation. In fact, the standard for both the spoken and written language is remarkably low. This is because Indonesians usually converse in their own local language (such as Javanese, Sundanese, Balinese, Chinese dialects, etc).

Due to the influence of the local dialects, the official language's pronunciation sounds different in different parts of the country. In addition, in the cities, people speak Indonesian with slang. All these various pronunciations and informalities cause a lot of inconveniences and problems during the database preparation process, which will be mentioned in following sections.

The set of Indonesian phonemes modified from that mentioned in Dessi's paper [28] is presented below. A full phoneme set contains 32 phonetic symbols (consist of 11 vowels which include diphthongs and semi-vowels, and 21 consonants) and one silent symbol in total. As can be seen from the table, vowel "e" actually has two phonemes which also gives two sounds, i.e. "e" and "E" which have sound of "e" in word "pen" and "property" respectively. However, for simplicity, both of them are referred as "e" in this Indonesian corpus.

Indonesian is considered to be one of the easiest languages to spell and pronounced. It is consistent in matching sounds to spellings with some exception. Below is some guidance in pronouncing Indonesian alphabets with respect to English.

There are 5 vowels, such as "a", "e", "i", "o", "u" and 3 diphthongs, such as "ai", "oi", "au"

- "a" is pronounced like the "a" in "father" in English
- "e" is a special vowel because there are two ways to pronounce this vowel.

"e" is pronounced like the "e" in "let" in English

"e" is pronounced like the "a" in "a house" in English

- "i" is pronounced like the "ee" in "seed" in English
- "o" is pronounced like the "o" in "open" in English
- "u" is pronounced like the "oo" in "moon" in English
- "ai" is pronounced like the "ie" in "die" in English
- "oi" is pronounced like the "oi" in "point" in English

- “au” is pronounced like the “ow” in “how” in English

Table 1 Indonesian Phoneme Set

Phonetic Category	Phoneme	Example	
		Word	Phoneme Sequence
Vowels	/a/	saya	/s a y a/
	/e/	enak	/e n a k/
	/E/	kEmana	/k E m a n a/
	/i/	ingin	/i n g i n/
	/o/	orang	/o r a n g/
	/u/	untuk	/u n t u k/
Diphthongs	/ai/	sungai	/s u n g a i/
	/au/	danau	/d a n a u/
	/oi/	amboi	/a m b o i/
Semi-vowels	/w/	wanita	/w a n i t a/
	/y/	saya	/s a y a/
Consonants			
Plosives	/b/	berapa	/b e r a p a/
	/p/	petani	/p e t a n i/
	/d/	dia	/d i a/
	/t/	teman	/t e m a n/
	/g/	giat	/g i a t/
	/k/	kamu	/k a m u/
	/kh/	khairul	/k h a i r u l/
Affricates	/j/	juga	/j u g a/
	/c/	cinta	/c i n t a/
Fricatives	/v/	video	/v i d e o/
	/f/	maaf	/m a a f/
	/z/	jenazah	/j e n a z a h/
	/s/	saya	/s a y a/
	/sy/	syahdu	/s y a h d u/
	/h/	hujan	/h u j a n/
Liquids	/r/	ramai	/r a m a i/
	/l/	lambat	/l a m b a t/
Nasals	/m/	mana	/m a n a/
	/n/	mana	/m a n a/
	/ny/	nyanyian	/n y a n y i a n/
	/ng/	lambang	/l a m b a n g/
Silence	/q/		

The consonants are mostly pronounced in the same way as the English pronunciation. However, there are some consonants with different pronunciation as English one, such as:

- “c” is pronounced like “ch” in “**ch**air” in English
- “p” is pronounced like “p” in “s**p**ot” in English
- “q” is pronounced similar to English “k”
- “r” is pronounced like the Spanish trilled
- “t” is pronounced like “t” in “s**t**op” in English

As mentioned, formal Indonesian is rarely used in daily conversations. The spoken Indonesian is modified from the formal Indonesian and it is used in informal writing as well to reflect the actual pronunciation.

For example:

1. In very informal spoken Indonesian, the diphthongs *ai* and *au* are pronounced as /e/ and /o/, *sampai* ([ENG] until) becomes *sampe*, *kalau* ([ENG] if) becomes *kalo*
2. In verbs, the prefix *me-* is often dropped, *mengambil* ([ENG] to take) becomes *ngambil*
3. If the basic word starts with consonant “c”, *mencuci* ([ENG] to wash), the prefix *me-* will be dropped and *ncuci* becomes *nyuci*
4. The suffixes *-kan* and *-i* are often replaced by *-in*, *mencarikan* ([ENG] help to find) becomes *nyariin*, *mengambilkan* ([ENG] help to take) becomes *ngambilin*
5. *kenapa* ([ENG] why) is pronounced as *kenapo* in Palembang and *kenape* in Jakarta, ect..

3.1.2 Recording and Transcription

All the recordings are selected from telephone conversations between the Indonesian native speakers. A Cenix digital recorder is connected between a phone and the phone socket during recording. One Indonesian native speaker will initiate a telephone conversation with another Indonesian native speaker and the two parties will talk over the phone. Then, the sentences spoken by the person at the receiving end will go through the Public Switched Telephone Network (PSTN) with 3.4kHz bandwidth, and then recorded by the digital recorder (with 12kHz bandwidth by default), while the conversation at the calling side is recorded without going through the telephone channel. The topics and the duration of the conversations differ from one to one.

After recording, the most tedious part of the database preparation is transcription. The main steps of transcribing are:

1. Cutting the conversation recordings into two sets of data. One set of data is calling side's talk, and the other set belongs to the person at the receiving end. Only the speech data collected from the receiving end are selected for the Indonesian corpus, since they are the narrow-band data that went through the PSTN.
2. Down-sampling the recordings to 8kHz. The default sampling rate of the recorder is set at 12kHz. Since the recordings went through the telephone channel, its effective bandwidth is limited to approximately 3.4kHz. Therefore, 8kHz sampling rate is good enough for this project according to Nyquist theorem.

- 3. Removal of the unsuitable portions for training, for example, when the background noise is too high, when the speech volume is too soft or laughter is combined with the talking, etc.
- 4. Segmenting the speech down to phrase level and label the phrase in text.

Table 2 Rules of Transcription Format

Scenario	Format	Examples
UPPER/lower case	use lowercase	
Abbreviations	no abbreviations	kallang road instead of kallang rd.
Letters	spell them out	d s o instead of dso
Numbers	spell them out	twenty one instead of 21
Stuttering	transcribe verbatim, use hyphen after word	i- i- i went to- to- to- the market
uh, ah, hm, ya, mm, huh, har	transcribe verbatim, don't use hyphens	uh huh this is uh good ya very good
different pronunciations	use [heard/intended]	[laylio/radio]
Informal pronunciations	use the word heard	cuma instead of cuman
missing syllables	put missing letters in []	i w[ent] to the market
foreign language	use [language-word]	[CHI-gou] means dog in Chinese
Laughter	use {L}	
laughter + speech	use {L-...}	{L-you are so funny}
Breathing	use {B}	
Noise	use {N}	
words can't be heard	use {UNK}	
start cut-off because of bad recording	use {CUT-...}	{CUT-hello} is john at home

During the transcription process, there are several rules and adjustment made to standardize the transcription data, such as words with prefix and or suffix are separated for simplification purposes. For example, *dibayar* ([ENG] is paid) becomes *di bayar*; *memenangkan* ([ENG] to win) becomes *me menang ka*. A full set of the rules is shown in Table 2.

The size of some recordings is very big, because it is recorded from a long conversation. However shorter sections are preferred by the recognition system. A shorter piece will result in better training as the transcription is not phone based but is phrase based, thereby reducing the difficulties of matching. Therefore, an additional splitting stage is required after transcription. A program for splitting file is developed. This program also makes the transcription file neat by deleting all the brackets in the original transcriptions as shown in Table 2.

3.1.3 Dictionary Preparation

Along with the data recording and transcribing, an e-dictionary for Indonesian is also needed. A required e-dictionary should contain a list of words and its pronunciation. Because there is no Indonesian e-dictionary available currently, the Indonesian dictionary together with the pronunciation was generated manually. A list of Indonesian words without pronunciation had been downloaded from web [31] as a starting point of the dictionary preparation.

Few entries of the dictionary were shown below as a preview.

a	a
a	e
a	e y
acar	a c a r
acara	a c a r a
acay	a c a y
accelerate	a k s e l e r e t
acceptable	a s e p t e b e l
accessor	a c c e s s o r
accu	a c c u
ada	a d a

During the training process, this Indonesian e-dictionary is updating all the time in two ways:

1. Multiple pronunciations: For example, word “ada” ([ENG] there is) is usually pronounced in two ways: “ada” and “ado”. Multiple pronunciations accommodate informal words or dialects that are spoken during the phone conversation. Pronunciation “ado” for word “ada” was a dialect pronunciation, spoken by people from West Sumatra. The dictionary updates to:

ada	a d a
ada	a d o

The entries are standardized that the transcription data contains all the formal pronunciations while the informal pronunciations are added to the dictionary list.

2. Foreign words: Some English words, Chinese words, as well as Japanese words may appear in the conversation, and they have to be added to the Indonesian dictionary with Indonesian phonemes. For example:

finalize	f ai n e l ai s
john	j o n

The original dictionary and the additional dictionary are then combined together using an HTK tool called HDMan. The tool creates a new dictionary that combines both the original and additional dictionary and deletes the multiple entries at the same time. The dictionary has a total of 50k entries to the present, and it will grow in size along with new telephone recordings.

3.2 Overview of Current Indonesian Database

The current Indonesian database is a corpus that consists of telephone conversations collected from 71 Indonesian native speakers (26 females and 45 males). All the phone conversations took place in different locations, some of which are plagued by high environmental noise. The total size of the speech corpus as yet is approximately 7 hours, all of which are manually transcribed. The structure and time duration of each speaker are given in Table 4.

All the speech files are put under a directory called IndonesianDB, and in its sub-root, “Train” and “Test” directories are created for training (manually transcribed data) and testing (untranscribed data) data respectively. All the 7 hours of transcribed data are put under “Train”, and they are grouped by speaker’s dialect region, which is referred to as the geographical area of Indonesia that they resided in during the childhood years. These geographical areas are in accordance to the recognized dialect regions in Indonesia. The respective dialect regions are specified in Table 3.

Table 3 Dialect Regions of Indonesia	
DR1	Sumatra Island
DR2	Java Island (except Jakarta)
DR3	Kalimantan Island
DR4	Sulawesi Island
DR5	Bali Island
DR6	Jakarta (Indonesian Capital)

The DR classification was done geographically, dividing Indonesia into 5 main islands, and the 6th region is Jakarta, Indonesia’s capital city. The people lived within each

classified region have similar dialect pronunciation. Inside the folder DRs, each person from that dialect region owns a folder by themselves, and these folders are named by 4 letters, FXXX or MXXX. The first letter stands for the gender of the speaker, female (F) or male (M). The next three letters are the first three letters of speakers' name. The *.wav (audio file of conversation) and *.phr (phrase level transcription text file) files are stored under each speaker correspondingly. A systematic file naming system is applied here. E.g., FXXX01 or MXXX02, where the last two digits are a serial number.

All the data in DR 1 to DR 6 are collected under actual telephone conversation environment and are not read from prepared prompts. Hence, the transcribing these data is a very tedious and time-consuming job.

In order to increase the robustness of the training data and also to speed up the transcribing process, an additional folder DR7 is created for speech read from script. The speaker was asked to read the text taken from an online Indonesian newspaper and it was recorded after going through the telephone channel. The phrases were chosen randomly from articles on several topics ranging from health to current affairs.

Table 4 Indonesian Speech Corpus Data

DR	REGION	INITIAL	DURATION (mins)	TOTAL DURATION PER REGION (mins)
1	Sumatra Island	FBET	1:32	100:41
		FFIA	14:15	
		FFID	2:01	
		FLIN	11:38	
		FLIS	7:49	
		FMAR	9:11	
		FNIT	1:31	
		FPAR	2:23	

		MALE	1:30	
		MAND	1:01	
		MAGU	8:50	
		MARI	1:00	
		MBOB	2:39	
		MERI	3:11	
		MFEN	2:19	
		MHAB	0:58	
		MHAR	1:36	
		MRIC	1:13	
		MRUS	8:14	
		MSAD	3:04	
		MTIM	1:36	
		MWAN	11:50	
		MYUD	1:20	
2	Java Island	FDEB	7:55	106:12
		FDEW	9:05	
		FELL	8:22	
		FJOH	5:30	
		FMEM	6:55	
		FMYR	10:10	
		MALB	1:50	
		MHAR	1:05	
		MHEN	1:01	
		MMOJ	7:50	
		MNIN	1:34	
		MPUT	15:56	
		MRON	4:25	
		MSOU	12:06	
		MVID	2:59	
		MWIB	9:29	
3	Kalimantan Island	FLIS	1:26	2:53
		FAND	1:27	
4	Sulawesi Island	MIVA	2:46	2:46
5	Bali Island	FPUT	20:25	63:11
		MAGU	8:26	
		MHEN	1:01	
		MMAG	16:32	
		MMAX	7:27	
		MWAY	11:20	
6	Jakarta Capital City	FADI	8:48	116:48
		FANI	1:26	

		FHES	4:51	
		FIME	1:19	
		FIRM	10:55	
		FKAT	1:34	
		FLIA	15:58	
		FSHI	1:08	
		MAND	7:09	
		MANT	4:58	
		MDHA	1:28	
		MEKA	11:03	
		MFER	7:16	
		MGIR	6:22	
		MHER	1:15	
		MISK	1:56	
		MRAY	6:12	
		MRON	1:11	
		MRUL	1:10	
		MSAL	4:53	
		MSEP	4:12	
		MWAR	5:55	
		MYOS	5:49	
7	Text Prepared	FELE	17:29	22:45
		FFIK	5:16	

3.3 Indonesian Recognition System

The whole recognition process for the Indonesian database is implemented by using Hidden Markov Model Toolkit (HTK). All HTK functions, which are shown in Figure 3, can be grouped into four common speech recognition stages. They are data preparation, training, testing and analysis as depicted in Figure 9.

Figure 9 Speech Recognition System Implemented by HTK

The requirement to build a set of HMMs is a set of speech data files and the corresponding transcriptions. The Indonesian database provides both of them. Before these speech data can be used in training, they must first be converted into a suitable parametric form, that is, the waveform has to be converted into a sequence of parameters, which is also called front end of a recognizer. A conventional 39 MFCC coefficients are used for the Indonesian recognizer. More details for training and testing stages will be

given in Section 3.3.1 and Section 3.3.2. The last stage of the whole recognition process, analysis, is to calculate the recognition rate of the recognizer using Equation 2.1.

3.3.1 Training

3.3.1.1 Define Topology of HMM

Figure 10 a Topology Model for Phoneme

After data preparation, the system defines a topology required for each HMM by writing a prototype definition. For large vocabulary, using HMM to represent individual words is impossible. So in this Indonesian recognizer, a phoneme is represented by a five states HMM in left to right topology.

The model shown above consists of 3 active states S_2 , S_3 and S_4 , associated with output probability distribution, while the first and last states S_1 and S_5 are non-emitting states. They are only used by HTK for implementation reasons. The observation functions b_i are Gaussian distributions and the transition probabilities are given by a_{ij} .

Total information needed for a single HMM is as follow:

1. Type of observation vector
2. number and width of each data stream

3. optional model duration parameter vector
4. number of states
5. for each emitting state and each stream
 - mixture component weights or discrete probabilities
 - if continuous density, then means and covariances
 - optional stream weight vector
 - optional duration parameter vector
6. transition matrix

3.3.1.2 Initialization of HMM parameters

Before the actual training process starts, the HMM parameters discussed in Section 3.3.1.1 must be properly initialized in order to allow a fast and precise convergence of the training algorithm. HTK offers two different initialization tools, namely HInit and HCompV. HInit is used when bootstrap data is available. HCompV that is used to create flat start monophones was chosen for this project. In this case, every state of the HMM is initialized to the same mean and variance vectors. These are computed globally on the whole training data set.

3.3.1.3 Multiple Gaussian Mixture Distributions

The philosophy of system construction in HTK is that HMMs should be refined incrementally. Thus, a typical progression is to start with a simple set of single Gaussian distribution for one state and then iteratively expand to multiple Gaussian mixture distributions. The tool HHed which is a HMM definition editor is used to increase the number of mixture components in specified distributions.

3.3.1.4 N-gram in Phoneme Level

Incorporation of knowledge of a certain language, in form of language model, is essential for large vocabulary speech recognition. A statistically motivated language model in phoneme level has been implemented in the training process. It is impossible to reliably estimate conditional word or phoneme probability for all sequence lengths for a given language. Hence, N-gram model is used in practice, which is only dependent on the preceding N-1 symbols.

The bigram in phoneme level has been tried in this project. HLSTATS provided by HTK is used to read in all the label files in phoneme level, build a table of bigram counts in memory, and then output a back-off bigram. The other tool HBuild can be used for converting bigram generated by HLSTATS into HTK lattice format.

3.3.1.5 Re-estimation process

Once an initial set of models has been created, the tool HERest is used to perform embedded training using the entire training set. HERest performs a single Baum-Welch re-estimation of the whole set of HMM phone models simultaneously. For each training utterance, the corresponding phone models are concatenated and then the forward-backward algorithm is used to accumulate the statistics of state occupation for example the means and variances, for each HMM in the sequence. When all of the training data has been processed, the accumulated statistics are used to compute re-estimates of the HMM parameters. HERest has to be applied after each time the HMM is modified, for example, after HHed increases the number of Gaussian Mixtures, HERest is called to re-estimate the new HMMs parameters.

3.3.2 Testing

The tool that is used in HTK to perform recognition is HVITE. There are a few types of input that HVITE can accept. They are a network describing the word sequences, a dictionary that defines the pronunciation of each word and a set of HMMs. The word network is first transformed into a phoneme network and the appropriate HMM definition is assigned to each phoneme instance. Recognition is then performed on a list of stored speech files or on a direct audio input, to find paths through network which have the highest log probability. HVITE uses the token passing algorithm to do Viterbi-based speech recognition. A token represents a partial path through the network extending from time 0 to time t . Tokens are propagated along the network transitions until an emitting HMM state during each time step. When there are multiple exits from a node, the token is copied so that all possible paths can be explored in parallel. A large network will have many nodes, and many possible token paths, and it is a heavy computation load for the computers. In order to reduce in computation, only tokens which have some chance of being amongst the eventual winners will be propagated. This process can be handled by HVITE as well, which is called pruning.

3.4 Tests and Results

3.4.1 DR Dependent Testing in Phoneme Level

DR dependent testing was carried out at first. The training source is all recordings in a single DR, and the testing data is the same as the training data. Recognition results in phoneme level of each DR in the current Indonesian Database are given in the following table.

Table 5 Phoneme Level Recognition Results for Each DR

Training and Testing Targets	Recognition Rate	Bigram Recognition Rate
DR1	61.33	52.74
DR2	55.34	53.30
DR3	91.18	88.54
DR4	92.84	90.05
DR5	61.36	58.48
DR6	56.80	54.04
DR7	88.31	89.92

The average recognition rate of unigram is around 72.45%, and the average recognition rate of bigram is around 69.58%, which is slightly worse than that of unigram. This is caused by the informal pronunciations in conversational speech. The bigram probability is calculated based on the standard pronunciations, while in the conversation, a lot of consonants are dropped, this will affect the accurate of the statistics. It is easy to find that, for DR7, the bigram result is better than the unigram result, which is because that all the speech in DR7 is read from script other than nature conversation that is closer to standard pronunciation. The results of DR3 and DR4 are extremely high, that is because that the data in DR3 and DR4 are very little that only contain two minutes and fifty three seconds and two minutes and forty six seconds conversation respectively.

3.4.2 DR Independent Testing in Phoneme Level

During DR independent test, the training target and the testing target are not from the same DR. Table 6 gives the results of one set of DR independent testing. In the test, DR2 is the training target while the other four DRs are the testing targets.

Table 6 Phoneme Level DR Independent Recognition Results

Training Targets	Testing Targets	Recognition Rate	Bigram Recognition Rate
DR2	DR1	51.19	46.62
DR2	DR3	51.74	45.51
DR2	DR4	50.08	46.22
DR2	DR5	48.10	46.31
DR2	DR6	47.35	44.58

The results show that, both the recognition rate for unigram and bigram of this set of tests are lower than that in the previous test set (DR dependent testing), which may be caused by the diversity in pronunciation of those DRs.

3.4.3 Overall Testing in Phoneme Level

Table 7 Overall Results in Phoneme Level

Training Targets	Testing Targets	Recognition Rate	Bigram Recognition Rate
DR1-7	DR1-7	53.84	48.49

The result of overall testing in phoneme level for the current Indonesian database is shown in Table 7.

3.5 Subjective Analysis over the Recognition Outputs

In the recognition process, recognition files called RECOUNT will be generated, which listed the recognition outputs of the testing speech by the recognizer in phoneme level. Then the recognizer would compare the RECOUNT file with the original transcription file of the corresponding speech to calculate the recognition rate by counting the number of recognized phonemes and the mistaken ones (insertions, deletions and substitutions).

However, from the observations, the recognition results are sometimes affected not only the recognizer but also by the database itself. Therefore, some subjective analyses have been done between the original transcriptions and the RECOUNT files.

3.5.1 Speed of the Speech

Table 8 Sample RECOUNT File for Speed of the Speech

ORIGINAL FILE

"*/dr1_FFIK_FFIK02MS10_te.lab"
walau

RECOUNT FILE

"D:/IndonesianRecognition/mfcc/dr1_FFIK_FFIK02MS10_te.rec"
200000 700000 w -250.871307
700000 1300000 l -381.164154
1300000 1800000 au -370.237305

The speed of the speech has an influence on the recognizer’s performance. If the speaker speaks too fast, the word may not be recognized correctly. From Table 8, the original file carried word “walau” (though), while, the RECOUNT file detected the word “w l au” instead, where “a” is missing. The reason might be the speed of the speech is too fast. When the speaker speaks very fast, the time duration for phoneme “a” is too short for the recognizer to catch. Therefore, it may be treated as noise or even silence during recognizing, which causes additional deletions during analysis. This situation occurs more frequently for speaker from certain dialect regions, e.g. Jakarta, who talks faster.

3.5.2 Multiple Dictionary Entries

The Indonesian dictionary for this recognizer is built to support multiple pronunciations. The placement of multiple pronunciations will affect the recognition outputs also. An

example is given in Table 9 shows an example for this case. Word “tiga” ([ENG] three) has two pronunciations and the dictionary lists both of them as follows:

tiga	t i g o
tiga	t i g a

The recognizer will capture the first entry in the dictionary by default. Therefore, “t i g o” appears in the RECOUNT file instead of “t i g a”, which causes additional substitutions during analyzing the recognition rate. This problem can be easily solved by applying force alignment during recognition.

Table 9 Sample RECOUNT File for Multiple Dictionary Entries

ORIGINAL FILE

"*/dr3_FLIS_FLIS01MS10_te.lab"
tiga

RECOUNT FILE

"D:/IndonesianRecognition/mfcc/dr3_FLIS_FLIS01MS10_te.rec"
300000 600000 t -326.871654
600000 1200000 i -265.165378
1200000 1600000 g -457.826450
1600000 2000000 o -305.471325

3.5.3 Informal Indonesian Pronunciations

Table 10 Sample RECOUNT File for Informal Indonesian Pronunciations

ORIGINAL FILE

"*/dr1_FFIK_FFIK02MS10_te.lab"
masih

RECOUNT FILE

"D:/IndonesianRecognition/mfcc/dr3_FFIK_FFIK02MS10_te.rec"
1800000 2200000 m -318.932465
2200000 2500000 a -223.518387
2500000 3000000 s -301.902405
3000000 3500000 i -418.649689

Another kind of mismatch occurred between the formal and informal Indonesian pronunciations. The transcription file always follows the formal pronunciation, while the recognition output file represents the exact pronunciation provided by the speaker, which are always the informal pronunciation in this conversational database. A sample RECOG file is given in Table 10.

The word “masih” ([ENG] still) has the standard pronunciation as “m a s i h”. Indonesian native speakers usually tend not to pronounce the “h” of the word when it is the last character. So the recognition output with pronunciation “m a s i” is correct to the speech file, while additional deletion is introduced as compared to the transcription file.

Chapter Four

Simultaneous Masking Modeling for ASR

Performance of an automatic speech recognition system drops dramatically in the presence of background noise unlike the human auditory system which is more adept at noisy speech recognition. Hence, a novel auditory modeling algorithm which is integrated with a feature extraction front-end for hidden Markov model (HMM) is proposed while preparing the database. The proposed algorithm simulates properties of the human auditory system (includes both simultaneous masking effect and temporal masking effect). This chapter will discuss the simultaneous masking modeling and the corresponding evaluation test results. Experiments carried out on AURORA 2 database, in comparison to ordinary MFCCs, show that the word recognition rate using the proposed feature extraction method has been effectively increased.

4.1 Introduction to Psychoacoustic

One of the biggest obstacles to widespread use of automatic speech recognition technology is robustness to those mismatches between the training data and testing data, which include environment noise, channel distortion and speaker variability. On the other hand, the human auditory system is extremely adept at all these mentioned situations [33, 34, and 35]. It therefore leads us to study the auditory model and apply it to an automatic speech recognition (ASR) system. In the following three chapters, a feature extraction front-end with noise-robust characteristics based on auditory evidence to improve performance of a speech recognition system is proposed.

Psychoacoustics is the science of quantifying the human perception of sounds, which comprises of relationship between sound pressure level and loudness, human response to different frequencies, and a number of masking effects [24]. Understanding of these neural responses has led to auditory modeling of these nonlinear perceptions by computational models. For example, as the most popular feature extraction algorithm for ASR, MFCC has employed an auditory-based warping of the frequency axis called critical-band, for modeling the frequency sensitivity of human hearing, and a first order pre-emphasis filter is applied in the time domain to boost the high frequencies [26]. My research is to integrate masking effects into the MFCC auditory model to make it more complete and to study the resulting effects when applied to automatic speech recognition.

Masking effects may be classified as simultaneous or temporal according to the occurrence of the signals. Masking effect between any two signals which occur at the same time is called simultaneous masking. Signals can also be masked by the preceding sound, called forward masking, or by the sound after it, called backward masking. Both forward masking and backward masking occur in the time domain, therefore they are known as temporal masking. Temporal masking and simultaneous masking are caused by the principal mechanism of neuronal signal processing both in time domain and frequency domain [36, 37, and 38].

4.2 Introduction to Simultaneous Masking

Simultaneous masking is more popular in speech enhancement than in speech recognition. A number of papers report that simultaneous masking works for speech enhancement [39,

40, and 41]. Lateral inhibition is one approach of simultaneous masking, which has been discovered through physiological, psychophysical and neurophysiological experiments [37]. Cheng's paper [42] shows that lateral inhibition can give a significant reduction of noise for noisy speech. An isolated word recognition technique based on a combination of instantaneous and dynamic features of the speech spectrum is proposed by Sadaoki [43]. Lu [44] suggests a feature extraction front-end that consists of both MFCC and lateral inhibition functions, but the lateral inhibition masker is a separate module after the MFCC as opposed to the proposed approach whereby the lateral inhibition masker is incorporated into the MFCC features.

Lateral inhibition is a sensory reception of a biological system by which neurons are able to determine more precisely the origin of a stimulus. This reception is caused by inter-connected neurons. When the skin is touched by an object, several neurons other than the exact one will be stimulated. In order to determine the origin of the stimulus, the biological system will tend to suppress the sideward neurons to make sure that only the center neurons will fire. In other words, the general function of lateral inhibition is to sharpen input changes.

Assume that a speech signal corrupted by noise is given by:

$$x(k) = s(k) + n(k) \quad (4.1)$$

where $s(k)$ is the clean speech, and $n(k)$ is white noise and uncorrelated with speech signal, then

$$P_x(\omega) = P_s(\omega) + P_n(\omega) = P_s(\omega) + \lambda \quad (4.2)$$

where $P_x(\omega)$, $P_s(\omega)$ and $P_n(\omega)$ are the power density spectra of $x(k)$, $s(k)$ and $n(k)$ respectively, and λ is a constant.

Let $M(\omega, \Omega)$ represent the lateral inhibition masker, where the two frequency indices indicate that the lateral inhibition phenomenon is frequency dependent. The lateral inhibition masker satisfies the following constraint:

$$\int_{-\infty}^{\infty} M(\omega, \Omega) d\omega = 0 \quad \forall \Omega \quad (4.3)$$

By applying the given lateral inhibition masker above to the noisy speech power spectral, the processed power spectrum is:

$$\begin{aligned} \hat{P}_x(\Omega) &= \int_{-\infty}^{\infty} P_x(\omega) M(\Omega - \omega, \Omega) d\omega \\ &= \int_{-\infty}^{\infty} P_s(\omega) M(\Omega - \omega, \Omega) d\omega + \int_{-\infty}^{\infty} P_n(\omega) M(\Omega - \omega, \Omega) d\omega \\ &= \hat{P}_s(\Omega) + \lambda \int_{-\infty}^{\infty} M(\Omega - \omega, \Omega) d\omega \\ &= \hat{P}_s(\Omega) \end{aligned} \quad (4.4)$$

which shows that the power spectrum of processed speech with white noise equals to that of the input clean speech theoretically. Therefore, the proposed algorithm not only sharpens the input signal on its spectrum but also removes noise.

Figure 11 shows the lateral inhibition effect of a stimulator (with a constant frequency F_c) on an inhibitor (with a constant intensity) as frequency varies. The frequency domain is

Mel scaled. When such a masker in the frequency domain is applied to speech signals, it helps to distinguish the central signal from others by enhancing the spectral peaks.

Figure 11 Sketch of Lateral Inhibition

4.3 Simultaneous Masking Modeling

The continuous lateral inhibition model based on Mel frequency scale requires a large amount of computation to be applied to a feature extractor [42]. For a practical and feasible system, a simplified lateral inhibition model for front-end feature extraction is proposed.

Although lateral inhibition is frequency dependent, the frequency discrimination of the human auditory system is approximately logarithmic in frequency. This means that by using Mel frequency scale, lateral inhibition can be approximated to a Mel frequency independent masker. Such an approximation therefore simplifies the lateral inhibition filter by reducing the number of variables. Figure 12 shows a simplified lateral inhibition

filter as a function of Mel frequency abscissa and the magnitude of its parameters, constrained by the zero mean condition.

Figure 12 Simplified Lateral Inhibition Model

According to Cheng's paper [42],

$$\Delta f = f_2 - f_1 = f_3 - f_2 = f_4 - f_3 = 1 \quad (4.5)$$

$$m_1 = -0.6 \quad (4.6)$$

$$m_2 = 1 \quad (4.7)$$

$$m_3 = -0.4 \quad (4.8)$$

is a set of optimal values, which corresponds to the asymmetrical characteristic of lateral inhibition in the frequency domain.

However, such a direct implementation over-attenuates some coefficients, leading to poorer recognition results. On the other hand, a combination of the original signals and the lateral inhibition outputs results in the best performance as shown:

$$\begin{aligned}
 \bar{P}_x^i(\omega) &= (1 - \beta)P_x^i(\omega) + \beta \sum_{j=-N}^N M_j P_x^{i+j}(\omega) \\
 &= (1 - \beta + \beta M_0)P_x^i(\omega) + \beta M_{-2}P_x^{i-2}(\omega) \\
 &\quad + \beta M_{-1}P_x^{i-1}(\omega) + \beta M_1P_x^{i+1}(\omega) + \beta M_2P_x^{i+2}(\omega)
 \end{aligned} \tag{4.9}$$

where $M_{-2} = m_1 = -0.6$, $M_0 = m_2 = 1$, $M_2 = m_3 = -0.4$ and $M_{-1} = M_1 = 0$.

After a series of experiments, the optimal value of $N = 2$ and $\beta = 0.1$ is fixed for the rest experiments. Therefore, the final lateral inhibition filter integrated into the feature extraction front-end is given below:

$$[M_{-2} \quad M_{-1} \quad M_0 \quad M_1 \quad M_2] = [-0.06 \quad 0 \quad 1 \quad 0 \quad -0.04] \tag{4.10}$$

and it is convolved with the Mel spectrum.

The outputs after lateral inhibition masking may have some negative values, which are meaningless. Hence a rectifier is used after the lateral inhibition masking to set all negative outputs to zero.

Lateral inhibition is very sensitive to spectral changes and it will enhance the peaks in both speech signals and noise signals. Noise signals which do not have fairly flat power spectra will also be enhanced like the speech signals thereby greatly degrading the recognition results. Hence, a method is needed to alleviate this effect and that is to employ a temporal integration in spectral domain:

$$\bar{P}_x^i(\omega) = \frac{1}{2N+1} \sum_{j=-N}^N Z_j P_x^{i+j}(\omega) \tag{4.11}$$

In the proposed front-end algorithm, the parameters are set as:

$$N = 2 \quad (4.12)$$

$$[Z_{-2} \ Z_{-1} \ Z_0 \ Z_1 \ Z_2] = [0.4 \ 1.3 \ 1.6 \ 1.3 \ 0.4] \quad (4.13)$$

which were reported in Cheng's paper[42].

4.4 Test and Results

4.4.1 Overview of the Testing Database

Performance of the proposed front-end is evaluated on the AURORA 2 database [32] which contains speech data sampled at 8kHz. There are two training conditions in AURORA 2. The clean training set in this database has no noise added and it consists of 8440 utterances recorded from 55 male and 55 female adults. 4004 utterances from 52 male and 52 female speakers are split equally into 4 subsets with 1001 utterances each, with all speakers being present into each subset. In the multi-condition training set, four types of noises have been added at various SNR levels.

In test set A, four types of noises (subway, babble, car and exhibition hall) are added to the four different clean data subsets at SNRs of -5dB to 20dB with 5dB step size. So there are $4 \times 7 \times 1001 = 28028$ utterances in this test set. In test set B, four different types of noises (restaurant, street, airport and train station) are added to the same subsets of clean data with the same SNR levels. In test set C, subway and street noises filtered by a MIRS (Modified Intermediate Reference System) filter which simulates the frequency characteristics of a telecommunication terminal are added to the subsets of clean data.

The data size of test set C is thus half that of test sets A and B, since there are only two types of noise added.

4.4.2 Front End for Baseline System

The VoiceBox Toolkit reference code for MFCC is used to calculate 13 cepstral coefficients including log energy term (logE). Furthermore, delta coefficients and acceleration coefficients are computed by using

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \quad (4.14)$$

where d_t is a delta coefficient at time t computed in terms of the corresponding static coefficients $c_{t-\Theta}$ to $c_{t+\Theta}$. Θ is set to two in the experiments.

4.4.3 Recognizer Back End

The same recognizer is used for both the proposed front-end feature extraction algorithm and the baseline system for a meaningful comparison. Each digit is modeled by a simple left-to-right 18 states (including two non-emitting states) HMM model, with 3 Gaussian mixtures per state. Two pause models are defined. One is "sil", which has 3 HMM states and models the pauses before and after each utterance. The other one is "sp", which is a single state model (tied with the middle state of "sil") and models the pauses among words.

4.4.4 Lateral Inhibition without Weighting

In the very beginning of the test process, the set of parameters reported by Cheng's paper [42] has been tried out, which is $[M_{-2} \ M_{-1} \ M_0 \ M_1 \ M_2] = [-0.6 \ 0 \ 1 \ 0 \ -0.4]$.

However the recognition result is not good, due to over-attenuation of some useful coefficients. The comparison result of test Set A under clean training condition is listed in Table 11 below, where MFCC is the baseline system, while MFCC+LI stands for MFCC coefficients with lateral inhibition. The recognition rate of each noise type is the average recognition rate of SNR 0 to SNR 20.

Table 11 Set A Recognition Rate (%) of Lateral Inhibition without Weighting

	Subway	Babble	Car	Exhibition
MFCC	76.12	72.74	66.99	70.44
MFCC+LI	67.77	56.96	56.56	67.79

4.4.5 Lateral Inhibition with Weighting

After realizing 100% lateral inhibition results in over-attenuation, a set of experiments with different proportion of lateral inhibition processed coefficients and original coefficients were carried out.

Table 12 Set A Recognition Rate (%) of Lateral Inhibition with Weighting

	Subway	Babble	Car	Exhibition
MFCC	76.12	72.74	66.99	70.44
MFCC+100% LI	67.77	56.96	56.56	67.79
MFCC+50% LI	78.26	72.33	72.95	72.51
MFCC+20% LI	78.38	75.52	73.48	72.97
MFCC+10% LI	78.86	75.52	73.55	73.00

The test results of Set A are listed in Table 12, where MFCC is the baseline system, while MFCC+100%LI stands for MFCC coefficients with lateral inhibition, and MFCC+50%LI stands for the new cepstral coefficients' magnitude are made up by 50% of baseline coefficients and 50% of MFCC with lateral inhibition, etc. It is very clear to see that the combination of 10% lateral inhibition processed coefficients and 90% of the original

coefficients gives the best recognition rate among all. Hence the resultant lateral inhibition masker is:

$$\begin{aligned} [M_{-2} \ M_{-1} \ M_0 \ M_1 \ M_2] &= 0.1 \times [-0.6 \ 0 \ 1 \ 0 \ -0.4] + 0.9 \times [0 \ 0 \ 1 \ 0 \ 0] \\ &= [-0.06 \ 0 \ 1 \ 0 \ -0.04] \end{aligned} \tag{4.15}$$

4.4.6 Symmetrical and Asymmetrical Lateral Inhibition

After limiting the portion of lateral inhibition to 10%, different structures of lateral inhibition was experimented with. The test results of Set A are listed in Table 13 below. Among the symmetrical weighting ($M_{-2} = -0.05, M_2 = -0.05$), higher weighting on the left ($M_{-2} = -0.06, M_2 = -0.04$), and higher weighting on the right ($M_{-2} = -0.04, M_2 = -0.06$), the one with higher weighting on the left gives the best recognition rate. This is consistent with the asymmetrical properties of lateral inhibition.

Table 13 Set A Recognition Rate (%) of Symmetrical and Asymmetrical Lateral Inhibition

	Subway	Babble	Car	Exhibition
MFCC	76.12	72.74	66.99	70.44
$M_{-2} = -0.04, M_2 = -0.06$	76.89	74.49	72.19	71.72
$M_{-2} = -0.05, M_2 = -0.05$	78.01	74.92	72.53	72.30
$M_{-2} = -0.06, M_2 = -0.04$	78.86	75.52	73.55	73.00

4.4.7 Lateral Inhibition with Temporal Integration

Lateral inhibition is very sensitive to spectral changes and it will enhance the peaks in both speech signals and noise signals. So temporal integration is introduced to smooth some peaks, especially for the noisy peaks. The test results of Set A are listed in Table 14 below, where MFCC is the baseline system, MFCC+LI stands for MFCC coefficients

with lateral inhibition, and MFCC+LI+TI refers to the new cepstral coefficients that are processed with lateral inhibition and temporal integration.

Table 14 Set A Recognition Rate (%) of Lateral Inhibition with Temporal Integration

	Subway	Babble	Car	Exhibition
MFCC	76.12	72.74	66.99	70.44
MFCC+LI	78.86	75.52	73.55	73.00
MFCC+LI+TI	76.22	76.46	74.38	73.60

The overall spectrum of noise types like babble, car and exhibition noise are quite flat, the enhancing effects of lateral inhibition on the background noise are almost cancelled out by temporal integration, therefore resulting in improvement in recognition rate. However, for the periodic subway noise, the noise peak is quite sharp in the spectral domain and it is consistently present in neighboring frames. Hence, the temporal integration average is more effective in smoothing the speech spectrum than this particular type of background noise, which led to slight degradation in the recognition rate. Generally speaking, lateral inhibition with temporal integration has a positive effect on the overall recognition rate.

Chapter Five

Temporal Masking Modeling for ASR

This chapter will discuss the forward masking modeling of temporal masking and the corresponding evaluation test results. Experiments carried out on the AURORA 2 database, in comparison to ordinary MFCCs, show that the word recognition rate using the proposed feature extraction method has been effectively increased.

5.1 Introduction to Temporal Masking

As mentioned in Section 4.1, temporal masking contains both forward masking and backward masking. Forward masking reveals that over short durations, the usable dynamic range of the human auditory system is dependent on the spectral characteristics of the previous stimuli. Forward masking data had been collected and modelled by authors as early as 1964 such as Plomp [45]. The series of later papers, most significantly, in 1982 by Jesteadt, Bacon and Lehmann presented the most complete data set [46] for pure tone forward masking. However, even that was incomplete since the longest probe delay measured was 40ms, which is short of the necessary duration for complete adaptation. Consequently in the Strop and Alwan's paper [49], the equations used to determine the amounts of masking as well as the model parameters are different from predecessors.

According to results of pure tone forward masking experiments from [46], a probe following a masker is less audible than a probe following silence. Therefore, the amount

of forward masking can be viewed as a consequence of auditory adaptation to the masker which can be expressed as:

$$P = M(1 - m)(1 - b^d)a^u \quad (5.1)$$

where u is duration between masker and signal, d is masker duration, M is masker level and m, a, b are frequency dependent constants.

Strop and Alwan [49] had conducted experiments on five different American English speakers to build a new set of data for the dynamic model. In defining the model parameters, the dynamic adaptation stages are implemented as an adapting additive offset to the log energy to the signal.

After adapting to the masker, it takes time to readjust and hence recovery time is necessitated before the relatively less intense probe become audible. The duration is also a function of the duration of the masker, thereby reflecting the time required for the auditory system to adapt completely to the masker (see Figure 14).

Models of adaptation have been successfully applied in ASR [47, 48, 49, and 50]. Adaptation and the popular RASTA techniques have obvious similarities to each other [51]. However, forward masking has a recovery step after adaptation as reported by Oxenham [52]. Strop and Alwan [49] have demonstrated forward masking and its application to ASR. Tchorz and Kollmeier [48] tested various adaptation parameters in ASR system. Recently a related adaptation model has been developed and evaluated based on AURORA 2 and AURORA 3 [47].

Figure 13 Average Forward Masking Data (circles) and Standard Deviation (error bars), together with the Model Line Fit (lines) as a Function of Masker Level across Five Octaves, with Probe Delay of 15, 30, 60 and 120 ms as Parameter [49]

Figure 14 Example of Recovery Process for Forward Masking [49]

As discussed above, most of the studies done have either been on simultaneous masking or temporal masking, while only some preliminary research in the completed auditory model has been done so far. In Park and Lee's paper [53], a Mexican-hat convolutional filter is used to simulate a lateral inhibition masker which neglects the frequency asymmetry of the masking effect, rendering it less accurate. Regarding temporal masking, only forward masking was implemented. In addition, they only used a simple nearest-neighbor classifier as the recognizer in the experiments. A simple recognizer can register a better percentage improvement in performance given its lower base recognition rate compared to a more sophisticated HMM recognizer which is inherently more robust to noise.

Backward masking has the same characteristic but the masker comes after the signal other than before it. The backward masking is less effective than the forward masking. Hence, in the following sections, more attention has been paid to forward masking aspect of temporal masking.

5.2 Forward Masking Modeling

5.2.1 Type I Forward Masking Modeling

This type of forward masking model is based on three equations only. The first is for characterizing recovery, the second equation, which takes into account masker duration, to characterize attack and the last is to calculate the probe threshold difference.

Equation 5.2 is a first order difference equation which leads to an exponential decay of the logarithmic distance to target.

$$P = M(1 - m)a^u \quad (5.2)$$

As the duration between masker and probe u increases, the amount of masking effect, P , imposed on the incoming probe, becomes a smaller value. This equation is used to determine the model parameters m as well as a .

Equation 5.3 assumes incomplete adaptation and describes the amount of masking as a function of attack parameter, b , and discrete time masker duration, nd , as well as probe delay nu .

$$P = M(1 - m)(1 - b^{nd})a^{nu} \quad (5.3)$$

The difference between Equation 5.2 and Equation 5.3 is due to the assumption that complete adaptation had already occurred. The attack parameter b was determined through the following probe threshold difference equation.

$$\Delta P = M(1 - m)b^{nd}a^{nu} \quad (5.4)$$

The model parameters (m , a , and b) discussed above varies with the frequency of the input signal. Fives sets of data corresponding to the frequencies at 250Hz, 500Hz, 1000Hz, 2000Hz and 4000Hz are given in Table 15 [49]. The importance of these constants is that their ratio approximates that which is derived from physiological data presented by Goldhor in 1985 [54].

Table 15 Model Parameters for Forward Masking

Freq. Hz	Slope m	a	b
250	0.19	0.864	0.474
500	0.20	0.854	0.510
1000	0.26	0.816	0.543
2000	0.29	0.851	0.525
4000	0.34	0.858	0.507

Table 16 Centre Frequencies for Mel Scaled Filter Banks

Filter No.	1	2	3	4
$f_c(\text{Hz})$	46.88	152.34	230.47	320.31
Filter No.	5	6	7	8
$f_c(\text{Hz})$	414.06	519.53	632.81	761.72
Filter No.	9	10	11	12
$f_c(\text{Hz})$	898.44	1050.78	1214.84	1394.53
Filter No.	13	14	15	16
$f_c(\text{Hz})$	1593.75	1808.59	2046.88	2304.69
Filter No.	17	18	19	20
$f_c(\text{Hz})$	2589.84	2902.34	3242.19	3617.19
Filter No.	21	22	23	24
$f_c(\text{Hz})$	>4000	>4000	>4000	>4000

The frequency limits for the filter bank ranges from 0Hz up till 8kHz for the evaluation experiments. For the Mel scaled filter bank, the centre frequencies are important as they

are used to calculate the model parameters for determining the dynamic model. The centre frequencies for 24 Mel filter banks over 0Hz to 8kHz are tabulated in Table 16. It is easy to tell from Table 16, only 20 filter banks are useful in the evaluation experiment due to the Nyquis theory.

The model parameters (a , b and m) for each filter bank are calculated depending on the centre frequency of each filter bank. Every filter bank will share a set of model parameters. Since only five sets of model parameters are available, linear interpolation is required to derive the model parameters for all the filter banks. The weights are determined using the distances between the centre frequency and the two neighbour frequencies and the corresponding parameters at these two frequencies.

$$V = \frac{w_1 \cdot v_1 + w_2 \cdot v_2}{w_1 + w_2} \quad (5.5)$$

$$w_1 = \frac{f_2 - f_c}{f_2 - f_1} \quad (5.6)$$

$$w_2 = \frac{f_c - f_1}{f_2 - f_1} \quad (5.7)$$

where f_1 and f_2 takes on either one of the 5 frequencies(250Hz, 500Hz, 1000Hz, 2000Hz and 4000Hz), and $f_2 > f_1$, thus, values of f_1 and f_2 are corresponding to one of five frequencies that just below centre frequency (f_1) and one of five frequencies that just below centre frequency (f_2). The values of v_1 and v_2 are the model parameters at f_1 and f_2 .

Table 17 Model Parameters for Each Mel Filter Bank

Filter No.	f_c	Slope m	a	b
1	46.88	0.19	0.864	0.474
2	152.34	0.19	0.864	0.474
3	230.47	0.19	0.864	0.474
4	320.31	0.1928	0.8668	0.4841
5	414.06	0.1966	0.8706	0.4976
6	519.53	0.2023	0.8525	0.5113
7	632.81	0.2159	0.8439	0.5188
8	761.72	0.2314	0.8341	0.5273
9	898.44	0.2478	0.8237	0.5363
10	1050.78	0.2615	0.8178	0.5421
11	1214.84	0.2664	0.8235	0.5391
12	1394.53	0.2718	0.8298	0.5359
13	1593.75	0.2778	0.8368	0.5323
14	1808.59	0.2843	0.8443	0.5284
15	2046.88	0.2942	0.8512	0.5216
16	2304.69	0.3006	0.8521	0.5193
17	2589.84	0.3077	0.8531	0.5167
18	2902.34	0.3156	0.8542	0.5139
19	3242.19	0.3241	0.8553	0.5108
20	3617.19	0.3334	0.8567	0.5074

For 4th to 20th filter bank with the centre frequencies between 250Hz to 4000Hz, linear interpolation is used to obtain the model parameters for each Mel filter bank. While for 1st to 3rd filter banks with centre frequencies below 250Hz, the parameters at 250Hz can just be used .

5.2.2 Type II Forward Masking Modeling

Other than the first type of forward masking modeling, a first order infinite impulse response (IIR) high pass filter can be used as the Type II forward masking model. The transfer function of the high pass filter is

$$H(z) = \frac{2f_s\tau - 2f_s\tau z^{-1}}{1 + 2f_s\tau + (1 - 2f_s\tau)z^{-1}} \quad (5.8)$$

The relationship between input and output can be written as

$$y(n) \cdot (1 + 2f_s\tau) = 2f_s\tau \cdot x(n) - 2f_s\tau \cdot x(n-1) - (1 - 2f_s\tau) \cdot y(n-1) \quad (5.9)$$

$$y(n) = \frac{2f_s\tau}{1 + 2f_s\tau} x(n) - \frac{2f_s\tau}{1 + 2f_s\tau} x(n-1) - \frac{1 - 2f_s\tau}{1 + 2f_s\tau} y(n-1) \quad (5.10)$$

where $f_s = 128\text{Hz}$, the MFCC frame rate in the evaluation experiments and time constant $\tau = 240\text{ ms}$.

The filter's memory is initialized to zero, and the first frame of logarithmic Mel spectra is subtracted from all the frames when applying this filter to avoid an initial transient.

5.3 Test and Results

The front end of baseline system is the same as that in Section 4.4.1. The recognizer back end for both baseline system and the tested system with forward masking modeling are the same as that mentioned in Section 4.4.2.

5.3.1 Results of Type I Modeling

By applying Type I forward masking modeling to AURORA 2 database, the average recognition rate is increased by 5%. The test results of Set A are listed in Table 18 below.

It can be calculated from Table 18 that Set A’s average improvement is as high as

$$\frac{(80.56\% - 76.12\%) + (78.57\% - 72.74\%) + (79.81\% - 66.99\%) + (79.46\% - 70.44\%)}{4} = 7.87\%$$

which is higher than the 5% average improvement over the whole database. The increase for car noise is the highest at 12.19%.

Table 18 Set A Recognition Rate (%) of Type I Forward Masking Modeling

	Subway	Babble	Car	Exhibition
MFCC	76.12	72.74	66.99	70.44
MFCC+Type I FM	80.56	78.57	79.18	79.46

5.3.2 Results of Type II Modeling

Type II modeling performance is slightly better than Type I modeling. The overall improvement of Type II on AURORA 2 is 6.35%, as shown in Table 19. Set A contributes the most improvements, while Set C contributes the least improvements. The test results of Set A are listed in Table 20 below.

Table 19 Overall Recognition Rate (%) of Type II Forward Masking Modeling

	Set A	Set B	Set C	Average
MFCC	71.57	72.1	69.06	70.85
MFCC+Type II FM	79.95	82.28	69.36	77.20

Table 20 Set A Recognition Rate (%) of Type II Forward Masking Modeling

	Subway	Babble	Car	Exhibition
MFCC	76.12	72.74	66.99	70.44
MFCC+Type II FM	80.86	79.46	79.97	79.94

Chapter Six

Integrated and 2D Psychoacoustic Modeling for ASR

In this chapter, a novel auditory modeling algorithm which is integrated with a feature extraction front-end for hidden Markov model (HMM) is proposed. The proposed algorithm simulates properties of the human auditory system and applies it to the speech recognition system in order to enhance its robustness. It integrates simultaneous masking, temporal masking and temporal integration effects into ordinary Mel frequency cepstral coefficients (MFCC) feature extraction algorithm for robust speech recognition. The proposed method sharpens the power spectrum of the signal in both the frequency domain and the time domain. Experiments carried out on AURORA 2 database, in comparison to ordinary MFCCs, show that the word recognition rate has been effectively increased. The result also suggests that the proposed algorithm offers a simple and computationally negligible method which is competitive with MFCCs with cepstral mean normalization (CMN).

6.1 Integrated Psychoacoustic Modeling

6.1.1 Proposed Feature Extraction Front End

The proposed front-end feature extractor is modified from the MFCC model provided by Voicebox Toolkit [55] by integrating the lateral inhibition masking, temporal integration and forward masking. The completed integration of proposed new MFCC model with both simultaneous masking (lateral inhibition) and temporal masking (forward masking)

is given on the right of Figure 15, compared to the traditional MFCC model shown on the left of Figure 15.

Figure 15 Comparison between Traditional MFCC and Proposed MFCC with Masking Effects

6.1.2 Cepstral Mean Normalization (CMN)

Cepstral mean normalization is also implemented on the baseline system for comparison to the proposed system, since CMN is a commonly used feature enhancement method for robust speech recognition [56, 57, and 58]. The Hidden Markov Model Toolkit (HTK) is used to calculate the cepstral mean normalization in this experiment. The mean of MFCCs across frames is calculated by

$$\bar{c}_k = \frac{1}{T} \sum_{t=1}^T c(t, k) \quad k = 1, 2, \dots, K \quad (6.1)$$

where T is the number of frames and K is the number of coefficients in a vector. Then $\bar{c}_k$ is subtracted from the feature and then divided by the standard deviation.

$$\hat{c}(t, k) = \frac{c(t, k) - \bar{c}_k}{\delta_k} \quad k = 1, 2, \dots, K \quad (6.2)$$

This technique is very effective in practice where it compensates for long-term spectral effects such as communication channel distortion.

6.2 Test and Results

The baseline system is the same as that in Section 4.4.1. The recognizer back end for both baseline system and tested system are the same as that mentioned in Section 4.4.2.

Table 21 presents the overall comparative experimental performance for the baseline MFCC, MFCC+CMN(Cepstral Mean Normalization) which is a commonly used feature enhancement method for robust speech recognition and the proposed system, MFCC+LI (Lateral Inhibition)+TI (Temporal Integration)+FM(Forward Masking). The performance metric used is word recognition rate. Results are captured for both the clean and multi-condition training conditions. For the latter, the result is averaged over the three noisy test sets with SNRs ranging from 0 to 20dB. The largest improvement occurs at clean training condition using MFCC+LI+TI+FM, with a relative improvement in word recognition rate of approximately 9% over the baseline MFCC. This improvement is much bigger than the 4% improvement over MFCC by MFCC+CMN. For the multi-training condition, MFCC+LI+TI+FM register a lesser but still significant improvement over the baseline MFCC and the result is better than that of MFCC+CMN.

Table 21 Overall Recognition Rate (%) for AURORA 2

Training Cond.	MFCC(39)	MFCC+CMN	MFCC+LI+TI+FM
clean	70.85	73.69	79.57
multi	87.74	89.55	89.25

**Table 22 Recognition Rate (%) of MFCC+LI+TI+FM VS. MFCC (39)
under Clean Training Condition**

SNR/dB	MFCC+LI+TI+FM				MFCC average
	Set A	Set B	Set C	average	
Clean	99.48	99.45	99.48	99.47	99.35
20	97.68	98.37	96.21	97.42	96.72
15	96.87	97.68	91.67	95.41	93.20
10	90.67	93.40	80.42	88.16	80.37
5	74.08	78.59	60.98	71.22	55.47
0	47.97	52.16	36.77	45.63	28.51
-5	25.44	26.93	20.64	24.34	13.84
avg 0-20	81.45	84.04	73.21	79.57	70.85

The detailed comparison recognition results of MFCC baseline and MFCC+LI+TI+FM under clean training condition are given in Table 22 and the comparison results under multi-training condition are given in Table 23. The results show that the proposed algorithm yields better recognition rate for all SNR levels. It can be seen that it works very well for input signals with mid-range SNR, i.e. 0dB to 15dB, which is the most commonly encountered SNR range in actual environments. For example, in Table 22, the proposed algorithm yields a 5% improvement at 10dB, 16% improvement at 5dB and 17% improvement at 0dB.

**Table 23 Recognition Rate (%) of MFCC+LI+TI+FM VS. MFCC (39)
under Multi Training Condition**

SNR/dB	MFCC+LI+TI+FM				MFCC average
	Set A	Set B	Set C	average	
Clean	99.12	99.14	99.07	99.11	99.03
20	98.32	98.62	98.13	98.36	98.16
15	97.95	98.22	97.28	97.82	97.65
10	96.23	96.93	94.35	95.84	95.45
5	90.11	91.13	85.48	89.03	87.49
0	66.97	71.10	57.60	65.22	59.96
-5	28.77	34.23	25.99	29.66	26.44
avg 0-20	89.92	91.20	86.64	89.25	87.74

Bigger improvements can be obtained for certain noise types as shown in Figure 16 and Figure 17, which show the word recognition rate under varying SNR for car noise and train station noise respectively. It is clear that the recognition rate is improved by 20% to 30% within the mid-range SNR, which is a significant improvement for speech recognition systems under noisy environment, hence vindicating the fact that the proposed enhancement to the speech features has made the traditional recognizer more robust to noise.

Figure 16 Recognition Rate (%) under Car Noise

Figure 17 Recognition Rate (%) under Train Station Noise

To provide further insight into the performance superiority of the full psychoacoustic modeling (MFCC+LI+TI+FM) as compared to partial psychoacoustic modeling, experiments are performed on MFCC+LI, MFCC+LI+TI and MFCC+FM. As mentioned in Section 4.3., the temporal integration (TI) is implemented as a accessory to lateral inhibition (LI), so the LI and TI are combined together in Figure 18 and Figure 19, illustrates the recognition results of MFCC+LI+TI and MFCC+FM under both clean and multi-training conditions. According to experiments results, under clean training condition, forward masking contribute more for Test Set A and Test Set B, while later inhibition performs better for Test Set C. It is very clear that both lateral inhibition (LI) and forward masking (FM) have contributed a part to the total improvement of the

proposed method (MFCC+LI+TI+FM). Without either of it, the amount of improvement will be reduced.

Figure 18 Superiority of MFCC+LI+TI+FM under Clean Training Condition

proposed method (MFCC+LI+TI+FM). Without either of it, the amount of improvement will be reduced.

Figure 18 Superiority of MFCC+LI+TI+FM under Clean Training Condition

Figure 19 Superiority of MFCC+LI+TI+FM under Multi Training Condition

6.3 2D Psychoacoustic Modeling

6.3.1 Introduction to 2D Psychoacoustic Modeling

As shown in Figure 15, the lateral inhibition and forward masking are two separate steps in proposed MFCC+LI+TI+FM. Lateral inhibition only takes care of the masking effects for each frame in frequency domain, i.e. the masker occurred at time t with frequency f , $M(t, f)$, will affect the signals occurring at the same time with frequencies $f + \Delta f$ and $f - \Delta f$, i.e., $S(t, f - \Delta f)$ and $S(t, f + \Delta f)$. While forward masking only takes care of masking effects on the signals with the same frequency as the masker, which occur after the masker in time domain, i.e., if a masker, $M(t, f)$, will only has masking effects over

the signal, $S(t + \Delta t, f)$. However, according to the neurons responding tests [33-37], a strong masker can mask a weaker signal of a close frequency that occurs later in time. Hence, a pioneering 2D psychoacoustic modeling which covers the masking effect over a 2D surface based on time and frequency is proposed. With this proposed algorithm, the masker, $M(t, f)$, will have masking effect on the signals, $S(t + \Delta t, f \pm \Delta f)$.

6.3.2 2D Psychoacoustic Modeling and Parameters

It is possible to infer from Chapter 4 and 5 the following observations:

- Both lateral inhibition and forward masking are essential for sharpening the spectral peaks.
- Although forward masking is frequency dependent, the variation is observed to be relatively small (Table 15).

By noting the facts listed above, a simplified 2D psychoacoustic algorithm without backward masking effects has been implemented by making the following assumptions compared to the integrated MFCC+LI+TI+FM:

- Simultaneous masking (lateral inhibition) and temporal masking (forward masking) are sharing the same set of parameters.
- Lateral inhibition is symmetrical in itself.
- Forward masking is frequency independent.
- The parameters that are not available can be linearly interpolated.

A proposed front-end feature extractor is modified from the MFCC model provided by Voicebox Toolkit [55] by integrating a 2D psychoacoustic model. The traditional MFCC

model is shown on the left of Figure 20. The integrated MFCC+LI+TI+FM, which includes the lateral inhibition masking, temporal integration and forward masking, is shown in the middle of Figure 20. The MFCC model with 2D psychoacoustic effects is shown on the right of Figure 20. It very clear from the comparison that MFCC model with 2D psychoacoustic effects has a great advantage of simplicity over the integrated MFCC+LI+TI+FM. MFCC model with 2D psychoacoustic effects has only one extra step when compared to the traditional MFCC model, and has three steps less than the integrated MFCC+LI+TI+FM model.

Figure 20 Comparisons of MFCC, Integrated MFCC+LI+TI+FM and 2D Psychoacoustic Models

However, some reasonable assumptions listed above are necessary for achieving this simplicity, which is a minor trade-off of recognition rate when compared to the integrated MFCC+LI+TI+FM.

Figure 21 1D Mexican-hat Filter

The parameters of 2D psychoacoustic modeling is developed based on the 1D Mexican-hat, shown in Figure 21, is modeled by the difference of Gaussian

$$w(n) = A \exp(-n^2 / \delta_1^2) - B \exp(-n^2 / \delta_2^2), \quad n = -L, \dots, 0, \dots, L \quad (6.3)$$

where $2L + 1$ is the length of the masking filter. The most crucial problem is to determine the shape of the filter by parameters of A , B , δ_1 and δ_2 . For the evaluation tests in this thesis, $L = 4$, $A = 1.858$, $B = 0.858$, $\delta_1 = 0.74$, and $\delta_2 = 3.6$.

Firstly, a set of parameter for a 1D filter is calculated from the Equation 6.3, $[-0.07 \quad -0.27 \quad -0.16 \quad 1 \quad -0.16 \quad -0.27 \quad -0.07]$. By assuming that lateral inhibition and forward masking can share the same set of parameters, the initial parameter set is shown in Table 24 (A). Secondly, the parameters in another direction is calculated by linear interpolation and listed in Table 24 (B). Lastly, due to the stronger total masking effects, the enhancement of the center signal is necessary. According to experiments, 20 is an optimal value for the center signal, refer to Table 24 (C).

Table 24 2D Psychoacoustic Modeling Parameters
(A)

-0.07	0	0	0
-0.27	0	0	0
-0.16	0	0	0
3	-0.16	-0.27	-0.07
-0.16	0	0	0
-0.27	0	0	0
-0.07	0	0	0

(B)

-0.07	-0.06	-0.03	0
-0.27	-0.225	-0.11	-0.03
-0.16	-0.215	-0.225	-0.06
3	-0.16	-0.27	-0.07
-0.16	-0.215	-0.225	-0.06
-0.27	-0.225	-0.11	-0.03
-0.07	-0.06	-0.03	0

(C)

-0.07	-0.06	-0.03	0
-0.27	-0.225	-0.11	-0.03
-0.16	-0.215	-0.225	-0.06
20	-0.16	-0.27	-0.07
-0.16	-0.215	-0.225	-0.06
-0.27	-0.225	-0.11	-0.03
-0.07	-0.06	-0.03	0

6.4 Test and Results

The front end of baseline system is the same as that in Section 4.4.1. The recognizer back end for both baseline system and the tested system with forward masking modeling are the same as that mentioned in Section 4.4.2.

The test results of 2D psychoacoustic modeling over the whole database under clean condition and multi-condition are shown in Table 25 and Table 26 respectively. The results show that the proposed 2D filter algorithm yields better recognition rate for all SNR levels. It can be seen that it works very well for input signals in the mid-range SNRs, i.e. 0dB to 15dB, which is the most commonly encountered SNR range in actual environments. The same can be said for the MFCC+LI+TI+FM algorithm. For example, in Table 25, the proposed algorithm yields a 4.86% improvement at 10dB, 5.52% improvement at 5dB and 3.02% improvement at 0dB.

The comparison result of between the 2D filter and MFCC+LI+TI+FM is shown in Table 27. This basic 2D filter managed to achieve a reasonably good improvement over the MFCC baseline system. However, the improvement is 1%-2% less than that contributed by MFCC+LI+TI+FM, which is caused by those assumptions made for 2D filter.

**Table 25 Recognition Rate (%) of 2D Filter VS. MFCC (39)
under Clean Training Condition**

SNR/dB	2D Filter				MFCC average
	Set A	Set B	Set C	average	
clean	99.39	99.39	99.36	99.38	99.35
20	97.63	98.03	96.42	97.36	96.72
15	96.16	96.81	92.24	95.07	93.20
10	87.26	88.81	79.63	85.23	80.37
5	61.36	64.72	56.88	60.99	55.47
0	29.98	34.4	30.20	31.53	28.51
-5	14.58	16.51	14.06	15.05	13.84
avg 0-20	74.48	80.36	75.79	76.88	70.85

**Table 26 Recognition Rate (%) of 2D Filter VS. MFCC (39)
under Multi Training Condition**

SNR/dB	2D Filter				MFCC average
	Set A	Set B	Set C	average	
clean	99.22	99.22	99.07	99.17	99.03
20	98.28	98.48	97.73	98.16	98.16
15	97.84	98.13	97.00	97.66	97.65
10	95.89	96.64	94.06	95.53	95.45
5	89.28	89.80	84.03	87.70	87.49
0	63.14	64.00	53.12	60.09	59.96
-5	26.23	29.02	22.14	25.80	26.44
avg 0-20	88.89	91.04	87.50	89.14	87.74

**Table 27 Recognition Rate (%) of 2D Filter VS. MFCC+LI+TI+FM
under Multi Training Condition**

	2D Filter	MFCC+LI+TI+FM
Clean Condition	76.88	79.57
Multi Condition	89.14	89.25

Chapter Seven

Conclusion and Recommendations

7.1 Conclusion

A telephony conversational Indonesian database with 7 hours speech is reported in this thesis. The processes of recording, transcribing and dictionary preparation are described in detail. A preliminary phoneme level recognizer for Indonesian language is also presented.

Investigations into improvements that could be made to enhance the robustness of currently available ASR techniques have been reported in this thesis. It is well known that one of the major obstacles to robust speech recognition is the mismatch between the training and testing data features, caused by environmental noise. All robust ASR techniques aim to reduce this mismatch by minimizing the effect of noise.

A novel feature extraction algorithm based on MFCC, which is the most widely used feature extraction technique in robust speech recognition, is proposed. The proposed algorithm is based on masking effects of human psychoacoustic phenomenon. Base on the evaluation tests carried out using AURORA 2, simultaneous masking (lateral inhibition) contributes 4.2% improvement of recognition rate under clean training condition over the whole database, and temporal masking (forward masking) contributes 6.35% improvement of recognition rate under clean training condition over the whole

database. The complete version which combines both lateral inhibition and forward masking can contribute 8.72% improvement of recognition rate under clean training condition over the whole database. The pioneering 2D psychoacoustic model manages to have a reasonably good result over the MFCC baseline system, especially for the mid-range SNRs.

7.2 Recommendations for Further Research

7.2.1 Integration of Backward Masking

As discussed in this thesis, temporal integration consists both forward masking and backward masking. Backward masking is not integrated in the proposed feature extraction front end. Since backward masking is less significant as compared to forward masking, few researchers work on it. Hence there is currently no available data for backward masking for its effects to be properly modeled.

7.2.2 Improvement of the 2D Psychoacoustic Modeling

As discussed in Section 6.3, under certain assumptions, the proposed 2D psychoacoustic modeling is relatively simple, which is not completely consistent with the integration MFCC+LI+TI+FM modeling as mentioned in Section 6.1. The performance of this proposed simple 2D modeling is also not as good as that of the integration version (please refer to Section 6.2 and Section 6.4). Therefore, future work should be carried out to improve on the accuracy of the 2D psychoacoustic modeling by reducing those ideal assumptions.

7.2.3 Further Testing on more Databases

The other problem encountered in robust speech recognition is the inconsistency of algorithm performance across different databases. Even within the same database, some speech features work better under certain types of noise as compared to the others. One explanation for this inconsistency is that the varied nature of the databases, e.g., the nature of the utterance, the property of the background noise and the complexity of the sentences. To further test the robustness of the proposed techniques, more extensive testing on more speech corpus is recommended. The aim of future work should be a universal feature extraction approach which has stable performance, regardless of the databases and noise types.

7.2.4 Application of the Auditory Model to Indonesian Recognizer

The preliminary implementation of integrating the auditory models discussed in Chapters 4, 5 and 6 on the Indonesian recognizer on the Indonesian database discussed in Chapter 3 has been done. However, the results are not highlighted in this thesis as the results are not likely to be correct. The Indonesian database is not professionally transcribed and it contains some errors. The transcription is also word based and not phone based. The author is not conversant in Bahasa Indonesian and finds it difficult to correct for the transcription errors. To perform a proper Indonesian word level recognition, a statistical language model is a necessity that is also unavailable at this point in time. Hence, this is an interesting work for future research.

Author's Publications

This chapter shows the list of publications arising from the research work reported in this thesis.

1. Xuewen Luo, I. Y. Soon, and C. K. Yeo, "An Auditory Model for Robust Speech Recognition", International Conference on Audio, Language and Image Processing (ICALIP), vol.2, pp. 1105-1109, 2008
2. Xuewen Luo, I. Y. Soon, C. K. Yeo and S. N. Koh, "An Application of Lateral Inhibition to Robust Speech Recognition", IASTED International Conference on Signal and Image Processing, 2008
3. Xuewen Luo, I. Y. Soon, C. K. Yeo and S. N. Koh, "A Completed Model of Masking Effects for Robust Speech Recognition", submitted in year 2008 to Neurocomputing.

Bibliography

1. Speech Recognition Market Investigation by Allied Business Intelligence (ABI),
<http://www.abiresearch.com/abiprdisplay.jsp?pressid=225>
2. Lawrence Rabiner, Challenges in Speech Recognition, Lecture Notes for Rutgers University and the University of California at Santa Barbara
3. Dessi Puji Lestari, Koji Iwano, Sadaoki Furui, "A Large Vocabulary Continuous Speech Recognition System for Indonesian Language", 15th Indonesian Scientific Conference in Japan Proceedings. ISSN: 1881-4034, 2006
4. Sakriani Sakti, Eka Kelana, Hammam Riza, Shinsuke Sakai, Konstantin Markov and Satoshi Nakamura, "Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project", Proc. Of Workshop on Technologies and Corpora for Asia-Pacific Speech Translation, pp.19-24, 2008
5. Garfinkel, History of Speech Recognition and Transcription Software, 1998,
<http://www.dragon-medical-transcription.com/historysoeechrecognition.html>
6. M. Masroor Ahmed, Abdul Manan Bin Ahmed, "Review and Challenges in Speech Recognition", ICCAS 2005
7. D. Pallett, J. Fiscus, W. Fisher, J. Garofolo, B. Lund, and M. Prysbocki. "1993 Benchmark Tests for the ARPA Spoken Language Program", In ARPA (ARP 94) pp. 49—74
8. J. Cohen, H. Gish, and J. Flanagan, "Switchboard---the second year", Technical Report /pub/caipworks2 at ftp.rutgers.edu, CAIP Summer Workshop in Speech Recognition: Frontiers in Speech Processing II, July 1994

9. Juang, Biing-Hwang, "On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition – A Unified View", AT&T Bell Laboratories Technical Journal 63 (7), pp.1213-1243, 1984
10. Bourouba, E.-H., Bedda, M., Djemili, R. "Isolated Words Recognition System Based on Hybrid Approach DTW/GHMM", Informatica (Ljubljana) 30(3), pp. 373-384, 2006
11. Bengio Yoshua, De Mori Renato, Flammia Giovanni, Kompe Ralf , "Global Optimization of a Neural Network – Hidden Markov Model Hybrid", IEEE Transactions on Neural Networks, 3(2), pp. 252-259, 1992
12. N. Morgan and H. Bourlard, "Continuous Speech Recognition, an Introduction to the Hybrid HMM/Connectionist Approach", IEEE Signal Process. Mag., vol. 12, no. 3, pp. 24-42, May 1995
13. Hidden Markov Model Toolkits Homepage, <http://htk.eng.cam.ac.uk/>
14. Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Jukia Odell, Dava Ollason, Dan Povey, Valtcho Valtchev and Phil Woodland, "The HTK Book (for HTK Version 3.3)", Cambridge University Engineering Department, April 2005
15. Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, 77(2), P. 257-286, February 1989
16. James Neil Sneddon, "The Indonesian language: Its history and role in modern society", UNSW Press, 2004

17. S.B. Davis and P. Mermelstain, "Comparison of Parametric Representations for Monosyllabic Word Recognition", IEEE Trans. ASSP, vol. 28, no.4, pp. 357-366, 1990
18. L. Rabiner and B-H. Juang, "Fundamentals of Speech Recognition", Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1993
19. H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis of Speech" Proc. JASA., pp. 1738-1752, 1990
20. L. Rabiner and R. Schafer, "Digital Processing of Speech Signals", Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978
21. S. Furui, "Digital Speech Processing, Synthesis, and Recognition", Marcel Dekker Inc., New York, 2001
22. F. Jelinek, "Statistical Methods for Speech Recognition", MIT press, London, England, 1997
23. Thomas F. Quatieri, "Discrete-Time Speech Signal Processing", Prentice Hall PTR, 2002
24. Ben Gold, and Nelson Morgan, "Speech and Audio Signal Processing", John Wiley & Sons, Inc., 2000
25. S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Transactions on Signal Processing, 29(2), pp 254-272, 1981
26. B. Milner, "A Comparison of Front-end Configurations for Robust Speech Recognition", ICASSP, 1, pp 797-800, 2002

27. Hanson, B. A. and Applebaum, T.H., "Robust Speaker-independent Word Recognition using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy speech", ICASSP, vol.2, pp. 857-860, April 1990
28. Dessi Puji Lestari, Koji Iwano, Sadaoki Furui, "A large vocabulary continuous speech recognition system for Indonesian language", Proc. 15th Indonesian Scientific Conference in Japan (ISA-Japan), Hiroshima, Japan, pp.17-22, 2006
29. Sakriani Sakti, Eka Kelana, Hammam Riza, Shinsuke Sakai, Konstantin Markov and Satoshi Nakamura, "Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project", Proc. Of Workshop on Technologies and Corpora for Asia-Pacific Speech Translation, pp.19-24, 2008
30. http://en.wikipedia.org/wiki/Indonesian_language
31. A website powered by TruAlfa dictionaries, <http://indocit.com/>
32. H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of speech recognition system under noisy conditions", Proc. ISCA ITRW ASR, pp. 17-21, 2000
33. J.B. Allen, "How do humans process and recognize speech?", IEEE Trans. on Speech and Audio Processing., vol. 2, pp. 567-577, October 1994
34. H. Hermansky, "Should recognizers have ear?", Speech Communication, vol. 25, pp. 3-27, August 1998
35. B. Kollmeier, "Auditory principles in speech processing: do computers need silicon ears?", proc. Eurospeech, pp. 5-8, September 2003

36. Shihab A. Shamma, "Speech processing in the auditory system i: The representation of speech sound in the responses of the auditory nerve", J. Acoust. Soc. Amer., vol. 78(5), pp. 1612-1621, November 1985
37. Shihab A. Shamma, "Speech processing in the auditory system ii: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve", J. Acoust. Soc. Amer., vol. 78(5), pp. 1622-1632, November 1985
38. M. N. Kvale and C. E. Schreiner, "Short term adaptation of auditory receptive field to dynamic stimuli", J. Neurophysiol, vol. 91, pp. 604-612, 2004
39. J. Chen, J. Benesty and Y. (A.) Huang, "On the optimal linear filtering techniques for noise reduction", Speech Communication, vol. 49(4), pp. 305-316, April 2007
40. Virage Nathalie, "Single channel speech enhancement based on masking properties of the human auditory system", IEEE Trans. on Speech and Audio Processing., vol. 7(2), pp. 126-137, March 1999
41. A. Akbari, Azirani, Le Bouquin, R. Jeannes and G. Faucon, "Optimizing speech enhancement by exploiting masking properties of the human ear", Proc. ICASSP, vol. 1, pp. 800-803, 1995
42. Yan Ming Cheng and Douglas O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence", IEEE Trans. on Signal Processing., vol. 39(9), pp. 1943-1954, September 1991
43. Sadaki Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. on Acoustics, Speech and Signal Processing., vol. 34(1), pp. 52-59, September 1986

44. Lu Xugang, Li Gang and Wang Lipo, "Lateral inhibition mechanism in computational auditory model and its application in robust speech recognition", Neural Networks for Signal Processing X, vol. 2, pp. 785-794, December 2000
45. R.Plomp, "Rate of decay of auditory sensation," J. Acoust. Soc. Amer., vol 36, pp. 277-282, Feb 1964
46. W. Jesteadt, S. Bacon, and J. Lehman, "Forward Masking as a function of frequency, masker level and signal delay", J. Acoust. Soc. Amer., vol. 106, pp. 950-962, April 1982
47. M. Holmberg, D. Gelbart and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing", IEEE Trans. on Audio, Speech and Language Processing., vol. 14, pp. 43-49, January 2006
48. J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automation speech recognition", J. Acoust. Soc. Amer., vol. 106, pp. 2040-2050, June 1999
49. Brian Strobe and Abeer Alwan, "A model of dynamic auditory perception and its application to robust word recognition", IEEE Trans. on Speech and Audio Processing., vol. 5(5), pp. 451-464, September 1997
50. K. Aikawa, H. Singer, H. Kawahara and Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition", IEEE Int'l Conf. on Acoustics, Speech and Signal Processing., vol. 2, pp. 668-671, April 1993
51. H. Hermansky and N. Morgan, "Rasta processing of speech", IEEE Trans. on Speech Audio Process., vol. 2(4), pp. 578-589, October 1994

52. Andrew. J. Oxenham, "Forward masking: adaptation or integration?", J. Acoust. Soc. Amer., vol. 109(2), pp. 732-741, February 2001
53. Ki-Young Park and Soo-Young Lee, "An engineering model of the masking for the noise-robust speech recognition", Neurocomputing, pp. 615-620, 2003
54. R.S Goldhor, "Representations of consonants in the peripheral auditory system: A modeling study of the correspondence between response properties and phonetic features," RLE Tech Rep 505, Mass Inst. Technol, Cambridge, 1985.
55. Mike Brookes, MFCC feature extraction in Matlab from Voicebox is available at <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>
56. C. Mokbel, D. Juvet and J. Monne, "Deconvolution of telephone line effects for speech recognition", Speech Communication, vol. 19, pp. 185-196, September 1996
57. O. Viikki, D. Bye and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise", Proc. ICASSP, pp. 733-736, May 1998
58. W. Zhu and O'Shaughnessy, "Log-energy dynamic range normalization for robust speech recognition", Proc. ICASSP, pp. 245-248, 2005