การศึกษาคุณลักษณะสัทสัมพันธ์สำหรับการรู้จำเสียงพูดอินโดนีเซีย

นายแนซรูล เอฟเฟนดี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2549

A STUDY OF PROSODIC FEATURES FOR INDONESIAN SPEECH RECOGNITION

Mr. Nazrul Effendy

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Electrical Engineering
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2006

Thesis title:	A STUDY OF PROSODIC FEATURES FOR INDONESIAN SPEECH RECOGNITION
By:	Mr. Nazrul Effendy
Field of study:	Electrical Engineering
Thesis Advisor:	Associate Professor Dr. Somchai Jitapunkul

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctoral Degree

Lavansri .... Dean of the Faculty of Engineering
Professor Direk Lavansiri, Ph.D.

THESIS COMMITTEE:

Watt .... Chairman
(Associate Professor Dr. Watit Benjapolkul)

.... Thesis Advisor
(Associate Professor Dr. Somchai Jitapunkul)

.... Member
(Associate Professor Dr. Boonserm Kijsirikul)

Nisha .... Member
(Assistant Professor Dr. Nisachon Tangsangiumvisai)

.... Member
(Atiwong Suchato, Ph.D.)

แนซรูล เอฟเฟนดี, นาย: การศึกษาคุณลักษณะสัทสัมพันธ์สำหรับการรู้จำเสียงพูดอินโดนีเซีย. (A STUDY OF PROSODIC FEATURES FOR INDONESIAN SPEECH RECOGNITION) อ. ที่ปรึกษา : รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล 115 หน้า.

ข้อมูลชนิด Utterance ถูกใช้ใน spoken dialogue system ระบบการรู้จำเสียงพูด และ translation machine โดยทั่วไปใน spoken dialogue system ผู้ใช้สามารถถามคำถามหรือให้ข้อมูลกับเครื่องได้ spoken dialogue system จึงควรจะสามารถรู้จำผู้ใช้ได้เพื่อให้สามารถตอบสนองได้อย่างถูกต้อง วิทยานิพนธ์ฉบับนี้นำเสนอ the automatic utterance type recognizer เพื่อให้สามารถแยกแยะ declarative questions ออกจาก statements ในภาษาอินโดนีเซียได้ เนื่องจาก Utterance ในเสียงพูดทั้งสองชนิดมีลักษณะด้านคำและลำดับที่เหมือนกัน และแตกต่างกันเฉพาะในด้าน intonations การแยกแยะจึงต้องใช้ทั้งตัวรู้จำคำ และตัวรู้จำ intonations

ตัวรู้จำชนิด Utterance ถูกออกแบบบนพื้นฐานของแบบจำลองฟูจิซากิ ตัวรู้จำดังกล่าวจะใช้ผลรวมของค่าพารามิเตอร์จากแบบจำลองฟูจิซากิในการรู้จำ Utterance ในเสียงพูดทั้งสองชนิด ประสิทธิภาพสูงสุดของแบบจำลองฟูจิซากิเมื่อนำมาใช้ในการรู้จำ Utterance ได้จากผลรวมของ ค่า Fraction Value เท่ากับ $F_b/100$ ค่าแอมพลิจูดของ accent command และขนาดของ last phrase command เป็นสัญญาณข้าวของระบบ neural อย่างไรก็ตามการดึงค่าพารามิเตอร์มาใช้งานของแบบจำลองฟูจิซากิมีความซับซ้อนมากเกินกว่าจะสามารถนำมาใช้ได้ในระบบการรู้จำเสียงพูดแบบอัตโนมัติ ดังนั้นตัวรู้จำชนิด Utterance จึงถูกพัฒนาโดยใช้ค่าสัมประสิทธิ์ polynomial ของคอนทัวร์ความทุ้มแหลมของเสียง (Pitch Contours) ของคำสุดท้ายในแต่ละประโยค

ตัวรู้จำชนิด Utterance แบบอัตโนมัติซึ่งใช้ การแผ่ขยาย polynomial ประกอบด้วย การดึงคอนทัวร์ความทุ้มแหลมของเสียง และมอดูลการแบ่ง Utterance แบบอัตโนมัติ คอนทัวร์ความทุ้มแหลมของเสียงของ Utterance ในเสียงพูดแต่ละชนิดจะถูกวิเคราะห์เพื่อศึกษาถึงคำสุดท้ายของแต่ละชนิด Utterance นอกจากนี้โมเดลเสียงอินโดนีเซียได้ถูกออกแบบขึ้นเพื่อให้สามารถสร้างมอดูลการแบ่ง Utterance แบบอัตโนมัติได้ ผลการประเมินระบบแสดงว่าการใช้ข้อมูลคำสุดท้าย และการแผ่ขยาย polynomial สามารถแยกแยะ declarative questions ออกจาก statements ในภาษาอินโดนีเซียได้อย่างมีประสิทธิภาพ

ภาควิชา วิศวกรรมไฟฟ้า ลายมือชื่อนิสิต Effend ....

สาขาวิชา วิศวกรรมไฟฟ้า ลายมือชื่ออาจารย์ที่ปรึกษา ....

ปีการศึกษา 2549

##46718540: MAJOR ELECTRICAL ENGINEERING

KEY WORD: UTTERANCE-TYPE RECOGNIZER / INDONESIAN / DECLARATIVE QUESTION / STATEMENT

NAZRUL EFFENDY: A STUDY OF PROSODIC FEATURES FOR INDONESIAN SPEECH RECOGNITION, THESIS ADVISOR: ASSOCIATE PROFESSOR DR. SOMCHAI JITAPUNKUL, 115pp

Utterance-type information has been used in spoken dialogue system, speech recognition system and translation machine. In a typical spoken dialogue system, a user can ask a question or give information to the system. In another side, the spoken dialogue system should be capable of recognizing its user intention to give the correct response to him/her. In this dissertation, the automatic utterance-type recognizer is proposed to distinguish declarative questions from statements in Indonesian speech. Since utterances in these two types have the same words with the same order and differ only in their intonations, their classification requires not only a word recognizer, but also an intonation recognizer.

At first, the utterance-type recognizer is designed based on Fujisaki model. The utterance-type recognizer uses a combination of the Fujisaki-model-parameters as the features to recognize the two utterance types. The best performance of the Fujisaki model based utterance-type recognizer is achieved using a combination of a fraction value of F_b: $F_b/100$, the amplitude of last accent command, and the magnitude of last phrase command as the input of the neural networks. However, the Fujisaki parameter extractor is too complicated to be implemented in an automatic recognition system. Therefore, the utterance-type recognizer is developed using the polynomial coefficients of the pitch contours of the sentence's final word.

The automatic utterance-type recognizer using polynomial expansion consists of a pitch contour extractor, normalizer, feature extractor, classifier, and an automatic utterance segmentation module. The pitch contour of each utterance type is analyzed to investigate the final word of the two utterance types. To create the automatic utterance segmentation module, an Indonesian acoustic model is designed. The evaluation confirms that the method using the final word and polynomial expansion is effective to distinguish declarative questions and statements in Indonesian speech.

Department: Electrical Engineering   Student's signature ……………………

Field of study: Electrical Engineering   Advisor's signature ……………………

Academic year: 2006

ACKNOWLEDGEMENTS

I would like to express my great gratitude and deep appreciation to my advisor, Associate Professor Dr. Somchai Jitapunkul for his supports, guidance and friendship during completing my dissertation. I would like to thank my thesis committee chairman, Associate Professor Dr. Watit Benjapolkul, my dissertation committee members, Associate Professor Dr. Boonserm Kijsirikul, Assistant Professor Dr. Nisachon Tangsangiumvisai, and Atiwong Suchato, Ph.D. for any encouragements and suggestions for completing the dissertation.

I would like to express my sincere gratitude to AUN/SEED-Net and JICA for granting me a full scholarship and financial support for my research and study in Chulalongkorn University and my short-term study in Tokyo Institute of Technology, Japan, without which my Ph.D. program at Chulalongkorn University would not be finished.

I would like to thank Professor Dr. Sadaoki Furui, Associate Professor Dr. Koichi Shinoda and Assistant Professor Koji Iwano who help me a lot during my short-term study in Japan and for discussion about my research. I thank many students in the digital signal processing research group, Chulalongkorn University and in Furui Laboratory, Tokyo Institute of Technology from whom I benefit greatly through stimulating discussions.

Finally, I would like to express my deepest gratefulness to my family, who have entirely supported, encouraged, and believed in me throughout my study.

Table of Contents

Page

Page

List of Tables

Page

Page

List of Figures

Page

CHAPTER I
INTRODUCTION

Since the last one decade, several researchers have used utterance-type information in spoken system such as spoken dialogue system, speech recognition system and translation system. In a typical spoken dialogue system, a user can ask a question or give information to the system. In another side, the system should be capable of recognizing its user intention to give the correct response to him/her (Carpenter and Carrol, 1999). Intonation a speaker used can be a cue to identify the type of an utterance. In an automatic speech recognition system, the utterance-type information is used to decrease the word error rate of the system (Wright, 1998; Wright et al., 1999; Adami et al., 2003; Grau et al., 2004; Chen and Hasegawa-Johnson, 2005). In a translation system, the utterance-type information is used to resolve ambiguities in translating the utterances (Wahlster et al., 1997).

This dissertation focuses on the recognition of the declarative questions from statements in Indonesian speech. The two utterance types can contain the same words with the same order (Gunlogson and Christine, 2001; Brown-Schmidt et al., 2006) and differ in intonational means depending on individual speakers. Some speakers utter an utterance with a similar pitch contour that can be a declarative question or a statement. Even listeners can be confused in recognizing the utterance type whether a speaker is asking a question or stating a statement in some Indonesian dialects. When a listener is confused in recognizing the type of an utterance, he/she usually will ask the speaker to repeat the utterance. Therefore, to distinguish a declarative question from a statement in Indonesian utterance, a spoken dialogue system needs not only a word recognizer but also an intonation recognizer.

The same sentence can be uttered using the same utterance type in various pitch contours. The pitch contours may differ from sentence to sentence and from speaker to speaker (Akagi and Ienaga, 1995; Kuwabara and Sagisaka, 1995; Furui, 2001; Effendy et al., 2004). A speaker can utter an Indonesian word with a different stressed syllable without changing its meaning (Samsuri, 1978; Halim, 1981, Odé, 1994; Odé and van Heuven, 1998). Consequently, the variation of the pitch contour increases. Therefore, the recognition of the Indonesian utterances-types based on the pitch contour is a difficult task.

Since the last one decade, many utterance-type recognition methods have been investigated for other languages. The methods use N-gram models, hidden Markov models, Naïve Bayes classifiers, Bayesian networks, multilayer perceptrons, decision trees, transformation-based learning, and memory-based learning (Reithinger and Maier, 1995; Lee et al., 1997; Wright, 1998; Ries, 1999; Samuel et al., 1999; Keizer et al., 2002; Levin et al., 2003; Grau et al., 2004).

Pitch contour, the major correlate of the intonation (Yan et al., 2003; Watson and Hughes, 2006), has been shown as having a close relationship to the utterance types for Dutch and German (Haan et al., 1997; Brickmann and Benzmüller, 1999). In Dutch, interrogativity differs from declarativity by higher pitch, both local and global. On a local level, interrogative utterances are found to have high onsets as well as high offset. Globally, interrogativity causes utterances to be realized in a higher and a narrower register. In German, for the average speaker, the final boundary tone, the F0 range, and the slope of the top-line can be used to distinguish the utterance types.

In this dissertation, the parameters of Fujisaki model are used as the features to represent the pitch contour to distinguish declarative questions from statements in Indonesian speech (Effendy et al., 2004). Fujisaki model is adapted for recognition system, which is originally designed for speech synthesis system (Fujisaki and Ohno, 1995; Higuchi et al., 1997; Aguero et al., 2004).

Since the algorithm to extract the parameters of Fujisaki model is too complicated to be implemented in an automatic system, the utterance-type recognizer is further developed using polynomial coefficients of the pitch contour of the sentence's final word.

1.1 The Indonesian language

Indonesian is a non-tonal language. The Indonesian language, so-called Bahasa Indonesia, is a unity language formed from hundreds of languages spoken in the Indonesian archipelago. It was coined by Indonesian nationalists in 1928 and became a symbol of national identity during the struggle for independence in 1945.

Compared to other languages, which have a high density of native speakers, Indonesian is spoken as a mother tongue by only 7% of the population, and more than 195 million people speak it as a second language with varying degrees of proficiency. Approximately, there are 300 ethnic groups living in 17,508 islands, speaking 365 native languages or no less than 669 dialects (Tan). At home, people speak their own language, such as Javanese, Sundanese or Balinese, though almost everybody has a good understanding of Indonesian as they learn it in school (Sakti et al., 2004).

The standard Indonesian language is continuously being developed and transformed to make it more suitable to the diverse needs of a modernizing society. Many words in the vocabulary reflect the historical influence of various foreign cultures that have passed through the archipelago. It has borrowed heavily from Indian Sanskrit, Chinese, Arabic, Portuguese, Dutch, and English. Although the earliest records in Malay inscriptions are syllable-based written in Arabic script, modern Indonesian is phonetic-based written in Roman script (Quinn). It uses only 26 letters as in the English/Dutch alphabet.

1.2 Declarative Question and Statement in Indonesian Speech

Utterance types are also called utterance classes, dialogue moves, dialogue acts, and speech acts (Fishel, 2006). Utterances are normally grouped into four types: question, statement, exclamation, and command. Furthermore, questions can be divided into two types: open-questions and yes-no-questions (van Heuven and van Zenten, 2005). An open question involves question words such as ‘where’, ‘how’, and ‘why’. A yes-no question differs from the former type because it allows only two possible responses, positive (‘yes’) or negative (‘no’).

In Indonesian, a yes-no question can be generated (1) by using the question indicator 'apa' with or without the interrogative suffix ‘-kah’, (2) by using the interrogative ‘-kah’, and (3) by intonation (Halim, 1981). The yes-no question made by intonation is called a declarative question (Gunlogson and Christine, 2001; van Heuven and van Zenten, 2005).

The meaning of an utterance in Indonesian speech does not depend only on the words used by a speaker, but also on the intonation expressed by the speaker.

Although words used in two sentences are the same and in the same order, when a speaker uses two different intonations, the meaning of the two sentences might be different (Verhaar, 1992).

For example,

Statement:

Andi sedang membaca buku
(Andi is reading a book)

Declarative Question:

Andi sedang membaca buku?
(Andi is reading a book?)

A statement may be perceived as a declarative question when the speaker uses interrogative intonation in the same sentence with the same word order.

Questions are generally signaled by a high final pitch and overall higher pitch (Eady and Cooper, 1986; Hirst and Cristo, 1998; van-Heuven et al., 1999). In English, the question involves not only local *F0* variations, but also more global patterns (Pierrehumbert, 1980; Ladd, 1996; Xu, 2005). Unlike those of English, the utterance type characteristics of Indonesian have not been investigated extensively; only few researchers have studied these. Two of them are Ebing and Stack.

Ebing (1997) modeled the pitch contours of a word in Indonesian utterance using IPO (Institute for Perspective Research) approach (Ebing, 1997). Her basic model consists of five basic forms. Later, she developed these five basic forms into eight forms. She used the model to synthesize the speech and then examined it perceptively with the assistance of several Indonesian native listeners. She used the model to make a synthetic speech. The perceptive testing shows that the synthetic speech was comparable with the synthetic speech using original pitch contours.

Stack (2005) investigated some sentences uttered by two Indonesian native speakers of Manado Malay dialect (Stack, 2005). The word order in the sentence is arranged into several different structures. Her work showed that although the utterances are only from two speakers of the same dialect, the intonation units (pitch contour) of each word in the utterances are different.

1.3 Objective of the Dissertation

The objective of this dissertation is to study an automatic utterance-type recognizer to distinguish two utterance types in Indonesian speech: declarative question and statement.

1.4 Scope of the Dissertation

This dissertation covers the study of the automatic utterance-type recognizer to distinguish declarative question from statement in Indonesian speech. Followings are the scopes of the dissertation:

a) The utterance-type recognizer of declarative question and statement in Indonesian speech is design based on Fujisaki model,
b) Then, the utterance-type recognizer is developed to be an automatic recognition system using the polynomial coefficients of the pitch contour of the sentence's final word.
c) The characteristics of the final words of declarative question and statement in Indonesian speech are investigated to be used in the design process of the automatic utterance-type recognizer.
d) An Indonesian acoustic model is developed to be employed in the automatic utterance-type recognizer as an automatic utterance segmentation module.

1.5 Dissertation Outline

The dissertation consists of seven chapters. Chapter 1 describes the introduction of the dissertation and elaborates the motivation, the Indonesian language, the declarative question and statement in Indonesian speech, objective of the dissertation, scope of the dissertation, and dissertation outline. Chapter 2 provides a concise introduction of basic techniques used in the utterance-type recognizer. They consist of neural networks and hidden Markov model. Chapter 3 elaborates the experiments of the utterance-type recognizer based on Fujisaki model. The performance of the utterance-type recognizer using four combinations of the parameters of Fujisaki model is evaluated. Chapter 4 describes the Indonesian acoustic model, which will be used as the automatic segmentation module in the automatic utterance-type recognizer. Chapter 5 describes the characteristics of the final words of declarative question and statement in Indonesian speech. They consist of the pitch contours, the pitch range, the maximum pitch and the speech rate of the final words. Chapter 6 describes the experiments on the automatic utterance-type recognizer using the polynomial coefficients of the pitch contours of the sentence final words. The performance of the automatic recognizer is investigated using larger speech data, which cover larger variation of intonation from larger number of speakers than the speech data used in the investigation of the utterance-type recognizer based on Fujisaki model. Chapter 7 concludes the dissertation and explains the contributions of the dissertation and the future research of the Indonesian utterance-type recognizer.

CHAPTER II

FUNDAMENTAL TECHNIQUES FOR THE DESIGN OF AN UTTERANCE-TYPE RECOGNIZER

This chapter describes some fundamental techniques used in this dissertation to design an utterance-type recognizer. They consist of neural networks and hidden Markov model.

2.1 Neural Networks

The term neural network originally referred to the biological neural network system. In recent years, however, it has often been used to express an artificial neural network implemented on an electronic device such as a computer, and in fact, such artificial networks have been considered to be one of the most promising technological concepts for developing information systems such as pattern recognizers and function estimators (Katagiri, 2000).

Neural networks are actually based on the biological neural system. They are comprised of many of the key features of the biological systems, such as distributed computation mechanism, adaptivity (trainability), nonlinearity, and simplicity in the node computation.

The neural network consists of many nodes (circle in Figure 2.1), each of which conceptually corresponds to a neuron cell in the real biological neural system, and connections (arrow in Figure 2.1), each of which conceptually corresponds to an axon in the neural system. The nodes are mutually connected.

The neural network is trained using the pairs of data N input / target data vector pairs $D = \{x_n, t_n\}$. Each two-dimensional input vector is presented to the input layer, and the output of each input node equals the corresponding component in the vector. Each hidden node computes the weighted sum of its inputs to form its scalar *net activation*, which denoted simply as *net.* The *net* is the inner product of the inputs with the weights at the hidden nodes.

$$net_j = \sum_{i=1}^{d} x_i w_{ji} + w_{j0} = \sum_{i=0}^{d} x_i w_{ji} \equiv \mathrm{w}_j^t \mathrm{x}\,, \tag{2.1}$$

where the subscript i indexes nodes in the input layer, j in the hidden; w_{ji} denotes the input-to-hidden layer weights at the hidden node j. Each hidden node emits an output that is a nonlinear function of its activation, $f\,(net_j)$, that is,

$$y_j = f\,(net_j)\,. \tag{2.2}$$

This $f(\cdot)$ is sometimes called the *activation function* or merely 'nonlinearity' of a node.

Each output node similarly computes its net activation based on the hidden node signals as

$$net_k = \sum_{j=1}^{n_H} y_j w_{kj} + w_{k0} = \sum_{j=0}^{n_H} y_j w_{kj} \equiv \mathrm{w}_k^t\ \mathrm{y}\,, \tag{2.3}$$

where the subscript k indexes nodes in the output layer and n_H denotes the number of hidden nodes.

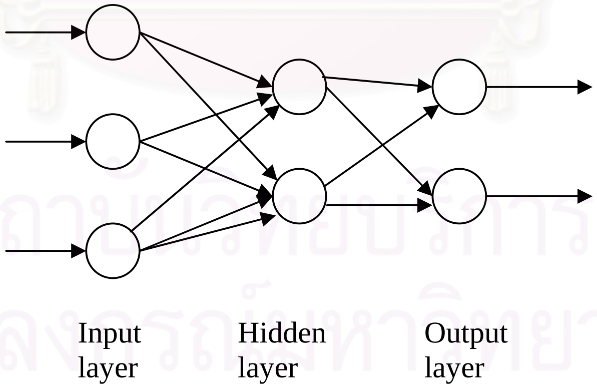


Figure 2.1 Multilayer neural network

2.1.1 Backpropagation Algorithm

Backpropagation is one of the simplest and most general methods for supervised training of multilayer neural networks. It is the natural extension of the LMS algorithm. Other methods may be faster or have other desirable properties, but few are more instructive.

The training error on a pattern is considered to be the sum over output nodes of the squared difference between the desired output t_k and the actual output z_k,

$$J(\mathrm{w}) = \frac{1}{2}\sum_{k=1}^{c}(t_k - z_k)^2 = \frac{1}{2}\|t - z\|^2 , \tag{2.4}$$

where t and z are the target and the network output vectors of length c and w represents all the weights in the network.

The backpropagation learning rule is based on gradient descent. The weights are initialized with random values, and then they are changed in a direction that will reduce the error:

$$\Delta\,\mathrm{w} = -\eta\frac{\partial J}{\partial\,\mathrm{w}} , \tag{2.5}$$

or in a component form:

$$\Delta w_{pq} = -\eta\frac{\partial J}{\partial w_{pq}} , \tag{2.6}$$

where η is the learning rate, and merely indicates the relative size of the changes in weight. The power of Equation (2.10) and (2.11) is in their simplicity. The weights are updated as

$$\mathrm{w}(m+1) = \mathrm{w}(m) + \Delta\,\mathrm{w}(m) , \tag{2.7}$$

where m indexes the particular pattern presentation.

Because the error is not explicitly dependent upon w_{jk}, the chain rule for the differential of J is:

$$\frac{\partial J}{\partial w_{kj}} = \frac{\partial J}{\partial net_k}\frac{\partial net_k}{\partial w_{kj}} = -\delta_k\frac{\partial net_k}{\partial w_{kj}} , \tag{2.8}$$

where the sensitivity of node *k* is defined to be

$$\delta_k = -\frac{\partial J}{\partial net_k}, \tag{2.9}$$

and describes how the overall error changes with the node’s net activation. Assuming that the activation function $f(\bullet)$ is differentiable, Equation (2.10) is differentiated and found that for such an output node, δ_k is simply

$$\delta_k = -\frac{\partial J}{\partial net_k} = -\frac{\partial J}{\partial z_k}\frac{\partial z_k}{\partial net_k} = (t_k - z_k)\, f'(net_k). \tag{2.10}$$

The last derivative in Equation (2.14) is

$$\frac{\partial net_k}{\partial w_{kj}} = y_j. \tag{2.11}$$

The weight update or learning rule for the hidden-to-output weights:

$$\Delta w_{kj} = \eta\delta_k y_j = \eta(t_k - z_k)\, f'(net_k)\, y_j. \tag{2.12}$$

Using the chain rule,

$$\frac{\partial J}{\partial w_{ji}} = \frac{\partial J}{\partial y_j}\frac{\partial y_j}{\partial net_j}\frac{\partial net_j}{\partial w_{ji}} \tag{2.13}$$

$$\begin{aligned}
\frac{\partial J}{\partial y_j} &= \frac{\partial}{\partial y_j}\left[\frac{1}{2}\sum_{k=1}^{c}(t_k - z_k)^2\right] \\
&= -\sum_{k=1}^{c}(t_k - z_k)\frac{\partial z_k}{\partial y_j} \\
&= -\sum_{k=1}^{c}(t_k - z_k)\frac{\partial z_k}{\partial net_k}\frac{\partial net_k}{\partial y_j}
\end{aligned}$$

$$\frac{\partial J}{\partial y_j} = -\sum_{k=1}^{c}(t_k - z_k)\, f'(net_k)\, w_{kj} \tag{2.14}$$

In analogy with Eq. (2.15), Eq (2.20) is used to define the sensitivity for a hidden node as

$$\delta_j = f'(net_j)\sum_{k=1}^{c} w_{kj}\delta_k. \tag{2.15}$$

The learning rule for the input-to-hidden weights is

$$\Delta w_{ji} = \eta x_i \delta_j = \eta \underbrace{\left[\sum_{k=1}^{c} w_{kj}\delta_k\right] f'(net_j)}_{\delta_j} x_i . \tag{2.16}$$

2.1.2 Neural Networks Application

Neural networks have been applied in many areas, such as pattern recognition, system modeling, digital signal processing and control engineering (Bishop, 1995; Effendy et al., 1998; Effendy et al., 2001; Effendy, 2002; Effendy et al., 2004). In speech technology, the neural networks have been applied for speech coding, speech recognition, speech signal processing, voice conversion, and speech enhancement (Katagiri, 2000).

2.2 Hidden Markov Model

The Hidden Markov Model (HMM) is a powerful statistical approach for the study of time series modeling with many of the classical probability distributions. The HMM approach provides a framework, which includes an automatic supervised training algorithm with mathematically proven convergence, the Baum-Welch algorithm. In addition, an efficient decoding scheme, the Viterbi algorithm, is incorporated in HMM. The underlying assumption of HMM is that the data samples can be well characterized as a parametric random process, and the parameters of the stochastic process can be estimated in a precise and well-defined framework.

Speech observation sequences corresponding to an acoustic event can be modeled by traversing an underlying sequence of connected states, each associated with an output distribution. The output distribution and the relative likelihood moving between states are estimated from a number of observation sequences of particular speech unit to be modeled. This is necessary to make speech recognition computationally tractable, and to ease the task of decoding a continuous waveform into a discrete set of symbols.

HMM has become one of the most successful statistical methods used in speech recognition, because of few assumptions need to be built into the models, and all model parameters can be efficiently estimated from the training data. Many successful speech recognition systems have employed the HMM approach as a major recognition part. Not only can the HMM be used in speech recognition, but it also can be applied in statistical language modeling, spoken language understanding, machine translation, and so on. In this dissertation, HMM was used as the engine of the automatic segmentation module of the automatic utterance-type recognizer.

This section briefly outlines theoretical framework of the HMM by explaining the definition of HMM. Then the essential algorithms needed to estimate the model parameters and decoding are described. All initial discussions are based on the discrete HMM. However, most of the discrete HMM concepts can be extended to the continuous HMM as described succeeding the discrete HMM.

2.2.1 Definition of the Hidden Markov Model

A natural extension to the Markov chain introduces a non-deterministic process that generates output observation symbols in any given state. Thus, the observation is a probabilistic function of the state. This new model is known as a hidden Markov model, which can be viewed as a double-embedded stochastic process with an underlying stochastic process or the state sequence not directly observable. The state sequence is hidden, and can only be observed through another set of observable stochastic processes.

A hidden Markov model is basically a Markov chain, where the output observation is a random variable generated according to the output probabilistic function associated with each state. A set of output probability distributions of each hidden state can be either discrete probability distributions or continuous probability density functions. To describe the HMM characteristics, the following HMM elements are defined.

1) The number of states in the model, N. Generally, the states are interconnected in such a way that any state can be reached from any other state. The individual states and the state at time t are denoted as $\boldsymbol{S} = \{S_1, S_2, ..., S_N\}$ and q_t respectively.

2) The number of distinct observation symbols per state, M. The observation symbols correspond to the physical output of the system being modeled. The individual symbols is denoted as $\mathbf{V} = \{V_1, V_2, ..., V_M\}$.

3) The state transition probability distribution, $\mathbf{A} = \{a_{ij}\}$ where

$$a_{ij} = P\left[q_{t+1} = S_j \middle| q_t = S_i\right], \qquad 1 \le i, j \le N . \tag{2.17}$$

4) The observation symbol probability distribution in state j, $\mathbf{B} = \{b_j(k)\}$, where

$$b_j(k) = P\left[V_k \text{ at } t \middle| q_t = S_j\right], \qquad 1 \le j \le N \; 1 \le k \le M . \tag{2.18}$$

5) The initial state distribution, $\boldsymbol{\pi} = \{\pi_i\}$, where

$$\pi_i = P\left[q_1 = S_i\right], \qquad 1 \le i \le N . \tag{2.19}$$

Since a_{ij}, $b_j(k)$, and π_i are all probabilities, they must satisfy the following properties:

$a_{ij} \ge 0$, $b_j(k) \ge 0$, $\pi_i \ge 0$ for all i, j, k

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{2.20}$$

$$\sum_{k=1}^{M} b_j(k) = 1 \tag{2.21}$$

$$\sum_{i=1}^{N} \pi_i = 1 \tag{2.22}$$

Given appropriate value of N, M, $\mathbf{A}$, $\mathbf{B}$, and $\boldsymbol{\pi}$, HMM can generate an observation sequence $\mathbf{O} = \boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_T$, where each observation $\boldsymbol{O}_t$ is one of the symbols from $\mathbf{V}$, and T is the number of observations in the sequence. A complete specification of an HMM requires two constant parameters, N and M, representing the total number of states and the size of observation symbols, and three sets of probability measures, $\mathbf{A}$, $\mathbf{B}$, and $\boldsymbol{\pi}$. For convenience, the compact notation is used to represent the complete parameter set of the model

$$\lambda = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}) \tag{2.23}$$

In the first-order hidden Markov model, there are two assumptions. The first is the Markov assumption for the Markov chain.

$$P\left(q_t \middle| q_1^{t-1}\right) = P\left(q_t \middle| q_{t-1}\right) \tag{2.24}$$

where q_1^{t-1} represents the state sequence $q_1, q_2, \ldots, q_{t-1}$. At each observation time t, a new state is entered based on the transitional probability, which only depends on the previous state. The transition may allow the process to remain in the previous state. The second is the output-independence assumption:

$$P\left(\boldsymbol{O}_t \middle| \boldsymbol{O}_1^{t-1}, q_1^t\right) = P\left(\boldsymbol{O}_t \middle| q_t\right). \tag{2.25}$$

The output-independence states that the probability that a particular symbol is emitted at time t depends only on the state q_t and is conditionally independent of the past observations. Although these assumptions severely limit the memory of the first-order HMM and may lead to model deficiency, in practice, they reduce the number of free parameters need to be estimated. Furthermore, these assumptions make evaluation, decoding, and learning feasible and efficient without significantly affecting the modeling capability.

2.2.2 Observation Density Functions

The observation density functions have to model the distribution of the feature vector for the different parts in data. These distributions are estimated from large amounts of training data. The most frequently distributions are listed below.

2.2.2.1 Discrete Density Functions

This type of density modeling requires that the multidimensional continuous observations be quantized into a number of symbols. Each state now has a discrete distribution that gives the probability of each symbol for the state. The discrete symbols are normally generated by a vector quantizer, which assigns a discrete symbol to each observation vector by choosing the nearest example from a small codebook of reference vector. This is implicitly dealt with the choice of distance metric for the clustering procedure in the vector quantization. The Euclidian distance measure, for instance, is used in the k -means clustering algorithm.

2.2.2.2 Continuous Density Functions

In this case, the observation probability distribution in state j, $b_j(\boldsymbol{O}_t)$, is a general parametric distribution of a predetermined form. The generalized method to continuous output density functions requires that the probability density functions be strictly log concave. The re-estimation algorithm can be extended to various types of elliptically symmetric density functions. The rationale of continuous density function is that the continuous observations can be directly modeled without quantization. However, the choice of different density functions to model a given observation largely depends on the characteristics of observations. In addition, a single continuous probability density function associated with each state is usually inadequate to model complicated observations. Therefore, finite mixture components are required.

2.2.3 Continuous Density Hidden Markov Model

If the observation does not come from a finite set, but from a continuous space, the discrete output distribution discuss in the previous sections can be extended to the continuous output probability density function. This implies that the vector quantization technique, which maps observation vectors from the continuous space to the discrete space, is no longer necessary. Consequently, the inherent error can be eliminated.

The generalized method to continuous output density functions can be applicable to the Gaussian, Poisson, and Gamma distributions but not to the Cauchy distribution. Furthermore, the estimation algorithm is expanded to cope with finite mixtures of strictly log concave and elliptically symmetric density functions. This section will discuss general re-estimation formulas for the continuous HMM, which is applicable to a wide variety of elliptically symmetric density functions.

2.2.3.1 Continuous Parameter Re-estimation

Using continuous probability density functions, the first candidate for a type of output distributions is the multivariate Gaussian, since

1) Gaussian mixture density functions can be used to approximate any continuous probability density functions in the sense of minimizing the error between two density functions,
2) By the central limit theorem, the distribution of the sum of a large number of independent random variables tends towards a Gaussian distribution,
3) The Gaussian distribution has the greatest entropy of any distribution with a given variance.

The most commonly used distribution is the continuous Gaussian density function defined as

$$\mathcal{N}(\boldsymbol{O};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}e^{-\frac{1}{2}(\boldsymbol{O}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{O}-\boldsymbol{\mu})} \tag{2.26}$$

where n is the dimensionality of the observation vector $\boldsymbol{O}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and the covariance matrix respectively. The advantage of normal distributions is that the parameters of Gaussian can be easily and reliably estimated from a large number of data. In order to obtain more accurate approximation, Gaussian mixtures are used. With enough components, such mixtures can approximate any density function with an arbitrary precision. The probability density of the multiple Gaussian mixtures is defined as

$$b_j(\boldsymbol{O}_t)=\sum_{m=1}^{M}c_{jm}\mathcal{N}(\boldsymbol{O}_t;\boldsymbol{\mu}_{jm},\boldsymbol{\Sigma}_{jm}) \tag{2.27}$$

where M is the number of mixture components and m is the mixture weight for the mixture component in state j. The mixture weights satisfy the stochastic constraint

$$\sum_{m=1}^{M}c_{jm}=1, \qquad 1\le j\le N \tag{2.28}$$

$$c_{jm}\ge 0, \qquad 1\le j\le N,\ 1\le m\le M \tag{2.29}$$

For the continuous probability density functions, the likelihood of an input observation is expressed as

$$P(\mathbf{O}|\lambda) = \sum_{all\ Q} P(\mathbf{O}, Q|\lambda) \tag{2.30}$$

$$= \sum_{all\ Q} P(Q|\lambda) P(\mathbf{O}|Q, \lambda) \tag{2.31}$$

An information-theoretic $\mathcal{Q}$-function, which is considered a function of $\bar{\lambda}$ in the maximization procedure, is applied to derive the re-estimation formula as

$$\mathcal{Q}(\lambda, \bar{\lambda}) = \frac{1}{P(\mathbf{O}|\lambda)} \sum_{all\ S} P(\mathbf{O}, Q|\lambda) \log P(\mathbf{O}, Q|\bar{\lambda}) \tag{2.32}$$

Using an auxiliary $\mathcal{Q}$-function, the re-estimated HMM parameters for the multimodal Gaussian distributions are

$$\bar{c}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t)}{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{jm}(t)} \tag{2.33}$$

$$\bar{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t) \cdot \boldsymbol{O}_t}{\sum_{t=1}^{T} \gamma_{jm}(t)} \tag{2.34}$$

$$\bar{\boldsymbol{\Sigma}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t) \cdot (\boldsymbol{O}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{O}_t - \boldsymbol{\mu}_{jm})'}{\sum_{t=1}^{T} \gamma_{jm}(t)} \tag{2.35}$$

where prime denotes vector transpose and $\gamma_{jm}(t)$ is the probability of being in state j at time t with the m^{th} mixture component for $\boldsymbol{O}_t$

$$\gamma_{jm}(t) = \left[\frac{\alpha_j(t)\beta_j(t)}{\sum_{j=1}^{N} \alpha_j(t)\beta_j(t)} \right] \left[\frac{c_{jm}\mathcal{N}(\boldsymbol{O}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\sum_{m=1}^{M} c_{jm}\mathcal{N}(\boldsymbol{O}; \boldsymbol{\mu}, \boldsymbol{\Sigma})} \right] \tag{2.36}$$

The re-estimation formula for a_{ij} is identical to the one used for discrete observation densities.

There are two possible options in the design of the mixtures. Either the Gaussian mixtures are state specific or they are shared (tied) between different states of the HMM. HMM with state specific Gaussian mixtures is called continuous density HMM. HMM that shares Gaussian mixtures among different states is called semi-continuous HMM or tied mixture HMM.

2.2.4 Hidden Markov Model for Speech Recognition

2.2.4.1 Composite Models for Continuous Speech Recognition

The parameter estimation and decoding techniques in the previous section are defined to apply to a single HMM mapped onto an isolated word. One of the advantages of the HMM approach is the ease with which it can be adapted to a continuous recognition environment. In order to extend to the continuous model, two modifications are made to the HMM structure; the addition of the entry and exit states to each model. The entry and exit states are defined as non-emitting states, which take Δt time to traverse, where Δt is negligibly small. Thus, the forward and backward probabilities that correspond to the entry and exit states are those at $t - \Delta t$ and $t + \Delta t$, where t is the time value at the immediately following or preceding state respectively. Therefore, the constraints are

$$a_{11} = 0 \text{ and } a_{Ni} = 0 \quad \forall i \tag{2.37}$$

which simply ensure that the entry and exit states can only be occupied for one transition. The other structural change is the addition of glue models. These models have only one emitting state, plus the entry and exit state, along with a non-zero entry to exit transition probability. These glue models are often call null or tee models (Young et al., 2002). A model with non-emitting entry and exit states is depicted in Figure 2.2 and a tee model is shown in Figure 2.3. Using tee models and non-emitting entry and exit states, a series of HMMs, with tee model between words, may be linearly combined into a single HMM for training purpose.

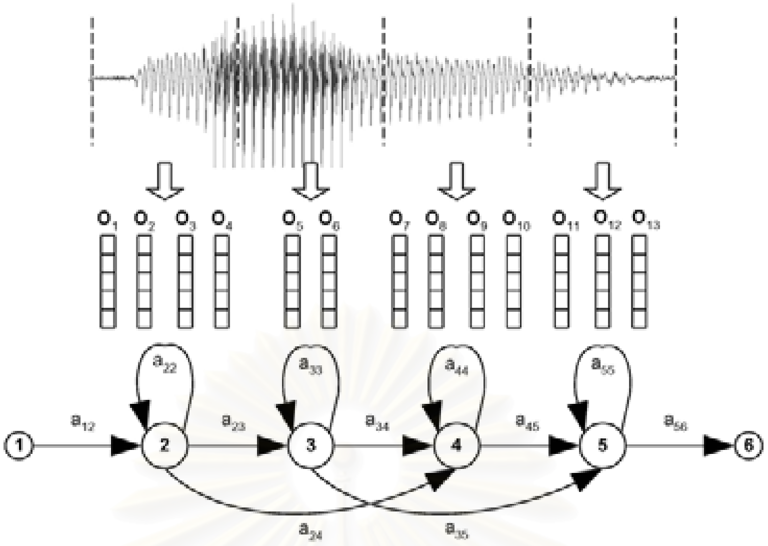


Figure 2.2 HMM with non-emitting entry and exit states

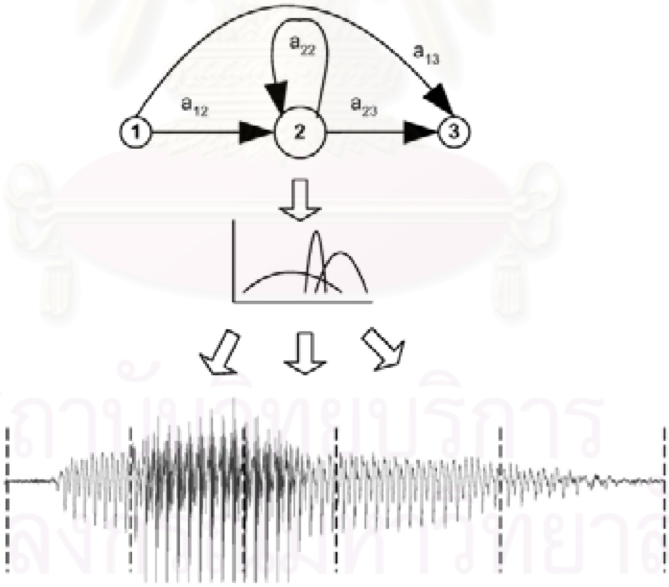


Figure 2.3 Tee model HMM

The modification required for the training formulas can be generated in a straightforward manner. The notation, a superscript q in parentheses representing the current model, is used as the notation that a training sentence model is represented by Q HMMs placed in sequence. The resulting forward and backward recurrent algorithms can be rewritten directly from the earlier definitions and new model structure. The forward equations are:

Initialization:

$$\alpha_1^{(q)}(1)=\begin{cases} 1 & q=1 \\ \alpha_1^{(q)}(1)a_{1N_q}^{(q-1)} & otherwise \end{cases} \tag{2.38}$$

$$\alpha_j^{(q)}(1)=\alpha_1^{(q)}(1)a_{1j}^{(q)}b_1^{(q)}(\boldsymbol{O}_t) \tag{2.39}$$

$$\alpha_{N_q}^{(q)}(1)=\sum_{i=2}^{N_q-1}\alpha_i^{(q)}(1)a_{iN_q}^{(q)} \tag{2.40}$$

Recursion:

$$\alpha_1^{(q)}(t)=\begin{cases} 0 & q=1 \\ \alpha_{N_{q-1}}^{(q-1)}(t-1)+\alpha_1^{(q-1)}(t)a_{1N_{q-1}}^{(q-1)} & otherwise \end{cases} \tag{2.41}$$

$$\alpha_j^{(q)}(t)=\left[\alpha_1^{(q)}(t)a_{1j}^{(q)}+\sum_{i=2}^{N_{q-1}}\alpha_i^{(q)}(t-1)a_{ij}^{(q)}\right]b_j^{(q)}(\boldsymbol{O}_t) \tag{2.42}$$

$$\alpha_{N_q}^{(q)}(t)=\sum_{i=2}^{N_q-1}\alpha_i^{(q)}(t)a_{iN_q}^{(q)} \tag{2.43}$$

The corresponding backward equations are:

Initialization:

$$\beta_{N_q}^{(q)}(T)=\begin{cases} 1 & q=1 \\ \beta_{N_{q+1}}^{(q+1)}(T)a_{1N_{q+1}}^{(q+1)} & otherwise \end{cases} \tag{2.44}$$

$$\beta_i^{(q)}(T)=\beta_{N_q}^{(q)}(T)a_{iN_q}^{(q)} \tag{2.45}$$

$$\beta_1^{(q)}(T) = \sum_{j=2}^{N_q-1} \beta_j^{(q)}(T) a_{1j}^{(q)} b_j^q(\boldsymbol{O}_T) \tag{2.46}$$

Recursion:

$$\beta_{N_q}^{(q)}(t) = \begin{cases} 0 & q=1 \\ \beta_1^{(q+1)}(t+1) + \beta_{N_{q+1}}^{(q+1)}(t) a_{1N_{q+1}}^{(q+1)} & otherwise \end{cases} \tag{2.47}$$

$$\beta_i^{(q)}(t) = \beta_{N_q}^{(q)}(t) a_{iN_q}^{(q)} + \sum_{j=2}^{N_{q-1}} \beta_j^{(q)}(t+1) a_{ij}^{(q)} b_j^{(q)}(\boldsymbol{O}_{t+1}) \tag{2.48}$$

$$\beta_1^{(q)}(t) = \sum_{j=2}^{N_q-1} \beta_j^{(q)}(t) a_{1j}^{(q)} b_j^{(q)}(\boldsymbol{O}_t) \tag{2.49}$$

The Baum-Welch re-estimation equations for transition probabilities are split into four categories:

1. internal transitions between emitting states,
2. transitions from the entry state into emitting states,
3. transition from emitting states into the exit state,
4. tee transitions from the entry state directly to the exit state, generally zero for non-tee models.

The equations are all similar to the original transition re-estimation formulas, with some primary differences above. The resulting formulas are:

$$a_{ij}'^{(q)} = \frac{\sum_{t=1}^{T-1} \alpha_i^{(q)}(t) a_{ij}^{(q)} b_j^{(q)}(\boldsymbol{O}_{t+1}) \beta_j^{(q)}(t+1)}{\sum_{t=1}^{T-1} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} \tag{2.50}$$

$$a_{1j}'^{(q)} = \frac{\sum_{t=1}^{T} \alpha_1^{(q)}(t) a_{1j}^{(q)} b_j^{(q)}(\boldsymbol{O}_t) \beta_j^{(q)}(t)}{\sum_{t=1}^{T} \alpha_1^{(q)}(t) \beta_1^{(q)}(t) + \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q)}(t)} \tag{2.51}$$

$$a_{iN_q}'^{(q)} = \frac{\sum_{t=1}^{T} \alpha_i^{(q)}(t) a_{iN_q}^{(q)} \beta_{N_q}^{(q)}(t)}{\sum_{t=1}^{T} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} \tag{2.52}$$

$$a'^{(q)}_{1N_q} = \frac{\sum_{t=1}^{T} \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q+1)}(t)}{\sum_{t=1}^{T} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} + \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q+1)}(t) \tag{2.53}$$

It can also be seen from examination of the last equation that the last model $q = Q$ in the state sequence cannot have a non-zero tee probability from the entry to exit state. This restriction is generally enforced for the initial model $q = 1$ as well, so that neither the beginning nor the end of an utterance sequence can be a tee model.

The underlying Baum-Welch equations for estimating output distributions do not change once the modifications have been made to the forward and backward probabilities.

2.2.4.2 Multiple Observation Sequence

In a complex large-vocabulary speech recognition system, there may be literally thousands of models representing context-dependent sub-word units or segmental sub-word units. One problem that arises when performing training operation is that the Baum-Welch equations discussed so far are designed to be computed on one training sentence at a time, which is likely to use only a handful of different models just once or twice each, resulting in a very small quantity of training data for each iteration and corresponding poor re-estimation.

A simple and accurate approach to solving is to treat the training sentences as a concatenated series of observation sequences assumed to be independent of each other. This concept leads to updating the parameters for each model only one time over the entire training set, where the new parameters are given by continuously summing the numerator and denominator terms of the re-estimation equations throughout training. In the transition probability re-estimations, a $\frac{1}{P_r}$ term, where P_r is the $\mathbf{P}(\mathbf{O}|\lambda)$ for the r^{th} sentence, is added to the numerator and denominator. The full set of re-estimation equations for the Gaussian mixture distributions with multiple observation sequences, including entry and exit states and tee models, is given below

$$a'^{(q)}_{ij} = \frac{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r-1} \alpha_i^{(q)}(t) a_{ij}^{(q)} b_j^{(q)}(\boldsymbol{O}_{t+1}) \beta_j^{(q)}(t+1)}{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r-1} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} \tag{2.54}$$

$$a'^{(q)}_{1j} = \frac{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_1^{(q)}(t) a_{1j}^{(q)} b_j^{(q)}(\boldsymbol{O}_t) \beta_j^{(q)}(t)}{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_1^{(q)}(t) \beta_1^{(q)}(t) + \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q)}(t)} \tag{2.55}$$

$$a'^{(q)}_{iN_q} = \frac{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_i^{(q)}(t) a_{iN_q}^{(q)} \beta_{N_q}^{(q)}(t)}{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} \tag{2.56}$$

$$a'^{(q)}_{1N_q} = \frac{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q+1)}(t)}{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \alpha_i^{(q)}(t) \beta_i^{(q)}(t)} + \alpha_1^{(q)}(t) a_{1N_q}^{(q)} \beta_1^{(q+1)}(t) \tag{2.57}$$

$$\gamma_{jm}^{(q)}(t) = \left[\frac{\alpha_j^{(q)}(t) \beta_j^{(q)}(t)}{P_r}\right]\left[\frac{c_{jm} b_{jm}^{(q)}(\boldsymbol{O}_t)}{b_j^{(q)}(\boldsymbol{O}_t)}\right] \tag{2.58}$$

$$c'^{(q)}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t)}{\sum_{r=1}^{R} \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{jm}^{(q)}(t)} \tag{2.59}$$

$$\boldsymbol{\mu}'^{(q)}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t) \cdot \boldsymbol{O}_t}{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t)} \tag{2.60}$$

$$\boldsymbol{\Sigma}'^{(q)}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t) \cdot (\boldsymbol{O}_t - \boldsymbol{\mu}_{jm})(\boldsymbol{O}_t - \boldsymbol{\mu}_{jm})'}{\sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{jm}^{(q)}(t)} \tag{2.61}$$

The implementation of these equations can be made with attention to some cancellations within the terms. In particular, the recursion for $\alpha_j^{(q)}(t)$ contains the term $b_j^{(q)}(\boldsymbol{O}_t)$ within it, which is also in the denominator of the formula for $\gamma_{jm}^{(q)}(t)$. The variable $U_j^{(q)}(t)$ is defined as

$$U_j^{(q)}(t) = \begin{cases} \alpha_1^{(q)}(t)a_{1j}^{(q)} & if\ t = 1 \\ \alpha_1^{(q)}(t)a_{1j}^{(q)} + \sum_{i=2}^{N_q-1} \alpha_1^{(q)}(t)a_{1N_q}^{(q)}\beta_1^{(q)}(t) & otherwise \end{cases} \tag{2.62}$$

to represent $\alpha_j^{(q)}(t)$ without $b_j^{(q)}(\boldsymbol{O}_t)$ term. The computation of this latter term is cancelled entirely, giving

$$\gamma_{jm}^{(q)}(t) = \frac{1}{P_r} U_j^{(q)}(t)\beta_j^{(q)}(t)c_{jm}b_{jm}^{(q)}(\boldsymbol{O}_t) \tag{2.63}$$

Similar modifications may be made to the distribution re-estimation equations for discrete probability densities so that composite models and multiple observation sequences can be considered, resulting in the equation

$$b_j'(\boldsymbol{O}_t) = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_j(t)}{\underset{s.t.\ s_j emits\ \boldsymbol{O}_t}{}\sum_{r=1}^{R}\sum_{t=1}^{T}\gamma_j(t)} \tag{2.67}$$

It should be noted that this formula has an identical form to the re-estimation equation for the mixture weights of Gaussian mixture distributions, if the mixture number m is treated as the index of the emitted observation. Thus, there is a direct correspondence between an M-mixture Gaussian distribution and a discrete distribution of M observation symbols.

2.3 Summary

This chapter provides concise information about fundamental techniques and methods in the utterance-type recognizer. It includes neural networks and Hidden Markov Model. Firstly, the neural network principle is described. Then, the Hidden Markov Model (HMM) is elaborated.

Neural networks are originally referred to the biological neural network system. They consist of nodes and connections between the nodes. The neural networks have been applied in many areas, such as pattern recognition, system modeling, digital signal processing and control engineering. In speech technology, they have been applied in speech coding, speech recognition, voice conversion and speech enhancement.

HMM is basically a Markov chain, where the output observation is a random variable generated according to the output probabilistic function associated with each state. It is a powerful statistical approach for the study of time series modeling with many of the classical probability distributions. It has become one of the most successful statistical methods used in speech recognition, because of few assumptions need to be built into the models, and all model parameters can be efficiently estimated from the training data.

CHAPTER III

UTTERANCE-TYPE RECOGNIZER OF THE DECLARATIVE QUESTION AND STATEMENT OF INDONESIAN SPEECH BASED ON FUJISAKI MODEL

In this research, Fujisaki model, which is originally designed for synthesis system (Fujisaki et al., 1996), is adapted for recognition system. Besides adapted for Indonesian intonation recognition system (Effendy et al., 2004; Effendy and Jitapunkul, 2006), Fujisaki model has been adapted for Thai tone recognition system (Potisuk et al., 1999; Ngarmchatetanarom et al., 2004).

Figure 3.1 shows the diagram block of an utterance-type recognizer of declarative question and statement of Indonesian speech based on Fujisaki model. The recognizer consists of a pitch extractor, Fujisaki model, a Fujisaki-model parameter extractor, and a classifier. The next sections describe the speech data, Fujisaki model and parts of the utterance-type recognizer.

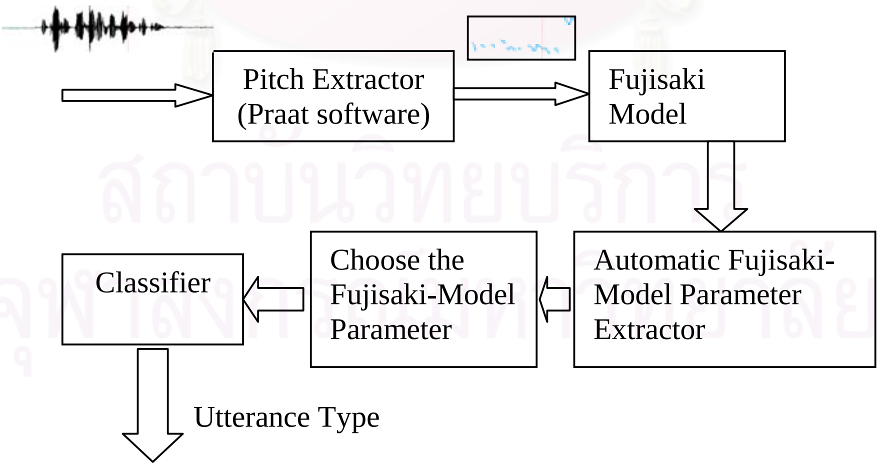


Figure 3.1 Utterance-type recognizer Based on Fujisaki model

3.1 Speech Data

To investigate the performance of the utterance-type recognizer, speech data consisting of 29 pairs of declarative questions and statements of Indonesian speech were used. The speech data are recorded from a native Indonesian male speaker. He was asked to utter the pairs of declarative questions and statements of Indonesian speech. The sentences are chosen from daily-life conversations among Indonesian speakers as listed in appendix A. The utterances are recorded at 16 kHz sampling rate, and 16-bit resolution in an office environment.

3.2 Fujisaki Model

Figure 3.2 shows a diagram of Fujisaki model. The model generates *F0* contours in the log *F* domain and is originally formulated for Japanese. It consists of phrase and accent commands. The phrase commands are assumed the impulse signals applied to the phrase control mechanism to generate the phrase components, while the accent commands are assumed the positive stepwise functions applied to the accent control mechanism to generate the accent components. Both mechanisms are assumed to be critically damped second-order linear systems, and the sum of their outputs: the phrase components and the accent components, is superimposed on a baseline value (*logFb*) to form an *F0* contour, as given by following equation (Fujisaki and Ohno, 1998):

$$\begin{aligned}\log F_0(t) = {} & \log F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) \\ & + \sum_{j=1}^{J} A_{aj} \left\{ G_a(t - T_{1j}) - G_a(t - T_{2j}) \right\}\end{aligned} \tag{3.1}$$

$$G_p(t) = \begin{cases} \alpha^2 t \cdot \exp(-\alpha t) & ,\text{for} \quad t \geq 0 \\ 0 & ,\text{for} \quad t < 0 \end{cases} \tag{3.2}$$

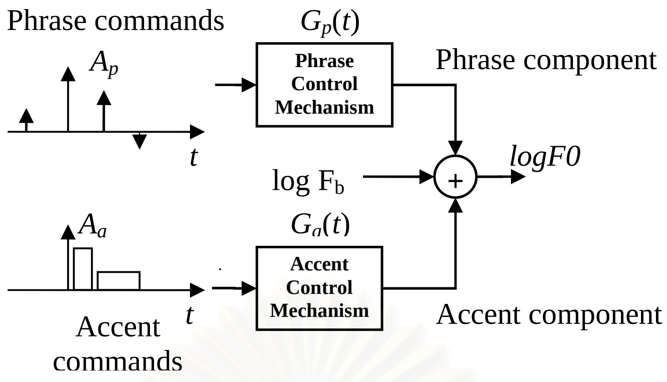


Figure 3.2 Fujisaki model (Fujisaki and Ohno, 1998)

$$G_a(t) = \begin{cases} \min\left[1-(1+\beta t)\cdot\exp(-\beta t),\gamma\right] & ,\text{for } t \geq 0 \\ 0 & ,\text{for } t < 0 \end{cases} \tag{3.3}$$

Where,

F_b : Baseline value of fundamental frequency

I : Number of phrase commands

J : Number of accent commands

A_{pi} : Magnitude of i^{th} phrase command

T_{0i} : Timing of i^{th} phrase command

A_{aj} : Amplitude of j^{th} accent command

T_{1j} : Onset of j^{th} accent command

T_{2j} : Offset of j^{th} accent command

α : Natural angular frequency of the phrase control mechanism

β : Natural angular frequency of the accent control mechanism

γ : Relative ceiling level of accent components

Equation (3.2) denotes the impulse response of the phrase control mechanism. The input signals to the phrase control mechanism are impulses, which are defined by their magnitude A_p and their onset time $T0$. α denotes the time constant of the phrase control mechanism and is assumed as being constant within an utterance. A_p determines the onset value of the F0 contour relative to *Fb*, unless an accent command is present.

$G_a(t)$ (Equation 3.3) denotes the step response of the accent control mechanism. The step-wise input signals to the accent control mechanism, the accent commands, are defined by their amplitude A_a, onset time $T1$ and offset time $T2$. β denotes the time constant of the accent control mechanism and is assumed as being constant in an utterance. The ceiling value γ (typically set to 0.9) of the accent control mechanism ensures that the accent component reaches its maximum in finite time. Hence, the change in $F0$ is in proportion to A_a.

The analysis of natural $F0$ contours is conducted by a method known as 'Analysis-by-Synthesis'. Basically, the number and parameters of input commands to the model is modified until an optimal approximation of the contour is yielded. When an arbitrary number of commands provided, any $F0$ contour can be approximated with unlimited accuracy. For this reason, constraints must be applied in order to ensure a linguistically meaningful interpretation of the analysis results. These constraints are language-specific and concern the relationship between linguistic units and structures (prosodic phrases and accents, for instance) and the phrase and accent commands.

3.2.1 Physiological Interpretation

In early works, Fujisaki (1971) explained the reason for formulating the model in the log F domain by the observation that contours by male and female speakers only look similar when plotted in the log F domain (Fujisaki et al., 1971). Fujisaki derived this property from the relationship between the tension T and elongation x of skeletal muscles (Fujisaki et al., 1981).

$$T = a(\exp(bx) - 1) \approx a\exp(bx) \quad \text{for exp } (bx) \text{ »1} \qquad (3.4)$$

The vibration frequency of elastic membranes varies in proportion to the square root of their tension.

$$F_0 = c_0\sqrt{T} \qquad (3.5)$$

Combining Equation (3.4) and (3.5) yields

$$\ln F0 = b/2x + \ln\left(\sqrt{a}c_0\right). \tag{3.6}$$

Hence ln F0 is in proportion to the elongation x plus a constant. The constant corresponds to the baseline value of the Fujisaki model, ln Fb. The oscillation frequency of the glottis (the vocalis muscle) is passively influenced by the cricothyroid muscle (CT) which changes the elongation x of the glottis. CT moves the thyroid cartilage relative to the cricoid cartilage, changing the length and thus the tension of the glottis. The movement has two degrees of freedom: 1) Rotation around the cricothyroid joint, 2) Translation of the thyroid against the cricoid.

3.2.2 Automatic Extraction of Fujisaki-Model parameters

Several researchers have investigated some methods of the automatic extraction of Fujisaki-model parameters. Among of them are Fujisaki, et.al. (1996), Rossi, et al (2002), Narusawa, et.al. (2002), Mixdorff (2000), Mixdorff, et.al. (2003), Silva and Netto (2004) (Fujisaki et al., 1996; Mixdorff, 2000; Narusawa et al., 2002; Rossi et al., 2002; Mixdorff et al., 2003; Silva and Netto, 2004).

The Fujisaki's model produces a particular F0 contour in the log F domain by superimposing three components: the phrase component, which corresponds to the phrase-wise slow overall declination line, the accent component made up by the faster movements in the F0 contour connected with accents and boundary tones, and Fb, a speaker-individual constant.

In order to separate the accent component from the phrase component and Fb, the spline contour is passed through a high-pass filter with a stop frequency at 0.5 Hz (Mixdorff, 2000). The output of the high-pass (henceforth called 'high frequency contour' or HFC) is subtracted from the spline contour yielding a 'low frequency contour' (LFC), containing the sum of phrase component and Fb. Hence, partial contours roughly corresponding to phrase and accent components are determined.

The initialization procedure makes use of the characteristics of phrase and accent command responses making up phrase and accent components, respectively. The phrase command response has its onset with the occurrence of an impulse-wise phrase command, rise to a maximum and then decays slowly according to the associated time constant α. Hence, in a sequence of phrase commands, the onset of a new command is characterized by a local minimum in the phrase component. Consequently, the LFC is searched for local minima, applying a minimum distance threshold of 1 sec between consecutive phrase commands. For initializing the magnitude value Ap assigned to each phrase command, the part of the LFC after the potential onset time T0 of a phrase command is searched for the next local maximum. Ap is then calculated in proportion to the frequency value found at this point. As responses of several phrase commands may add up in the phrase component, contributions of preceding commands must be taken into account when calculating Ap, which is reduced accordingly.

The accent command response is a smoothed square function rising from a value of 0 at T1 to a maximum which is sustained until the offset time T2 when it starts decaying. For initializing the appropriate number, onset times T1 and offset times T2 of accent commands, the HFC is searched for local minima, whose vicinity (± 100 msec) is scanned for even lower F0 values in order to avoid picking saddle points. Two subsequent local minima each are associated with a new accent command.

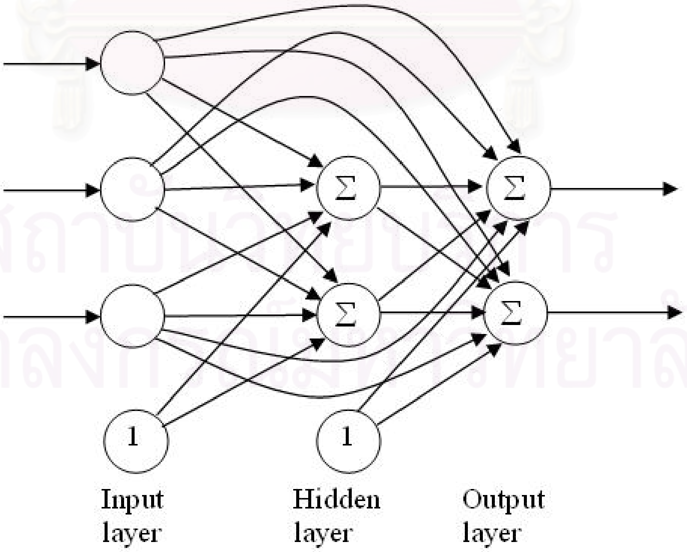


Figure 3.3 A specific fully connected feed forward neural network

3.3 Classifier

In this research, a specific fully connected feed forward neural network depicted in Figure 3.3 was used as a classifier to distinguish declarative questions from statements. The number of nodes in input layer can be changed to be arbitrary positive integer depending on the number of Fujisaki-model parameters used in the experiments. Both hidden layer and output layer consist of two nodes. The neural network is targeted to (0 1) for a statement and (1 0) for a declarative question.

3.4 Experimental Setup

All utterances in the speech data are extracted to produce the pitch contour using PRAAT software (©P.Boersma) (Boersma and Weenink, 2004). After the extraction of the pitch contour, Fujisaki-model parameters are extracted using an automatic Fujisaki parameter extractor proposed by Mixdorff and his colleagues (Mixdorff et al., 2003). Although, it is specifically designed for German, the automatic Fujisaki parameter extractor was adopted for the Indonesian language.

Twenty-nine utterances are used as the training set of neural network and fifty-eight utterances consisting also the training files are used as the testing set of the neural network. Three parameters of Fujisaki's model, the amplitude of last accent command (A_{aJ}), the magnitude of last phrase command ($A_{p\mathrm{I}}$); the baseline value of the fundamental frequency (F_b) are used to represent the pitch contour of each utterance type. Table 3-1 lists the example of the values of the Fujisaki-model parameters used in this dissertation as the input of the neural network. Four combinations of the Fujisaki-model parameters are created using the three parameters as illustrated in Table 3-2. The combinations are (1) the amplitude of last accent command (A_{aJ}), (2) A_{aJ} and the magnitude of last phrase command ($A_{p\mathrm{I}}$), (3) the baseline value of the fundamental frequency (F_b), A_{aJ} and $A_{p\mathrm{I},}$ and (4) a fraction of F_b: F_b/100, A_{aJ} and $A_{p\mathrm{I}}$.

Table 3.1 Example of Fujisaki-model parameters

No	F_b	A_{pI}	A_{aJ}	Utterance Type
1	107.07	0.2242	0.19	Statement
2	132.83	0.4012	0.60	Declarative Question
3	85.69	0.0924	0.19	Statement
4	119.13	0.1399	0.55	Declarative Question
5	120.02	0.3485	0.97	Statement
6	117.70	0.3149	0.51	Declarative Question
7	86.47	0.4673	0.24	Statement
8	108.37	0.1033	0.53	Declarative Question
9	97.81	0.3776	0.06	Statement
10	131.11	0.1877	0.25	Declarative Question
11	89.55	0.3473	0.16	Statement
12	132.96	0.0288	0.35	Declarative Question
13	106.02	0.4212	0.33	Statement
14	103.98	0.4814	0.53	Declarative Question
15	121.72	0.0518	0.27	Statement
16	116.54	0.043	0.53	Declarative Question
17	86.01	0.0025	0.47	Statement
18	98.88	0.2707	0.76	Declarative Question
19	98.18	0.489	0.20	Statement
20	106.67	0.2981	0.32	Declarative Question
21	89.46	0.42	-0.24	Statement
22	103.85	0.02	0.59	Declarative Question
23	97.47	0.22	0.17	Statement
24	98.64	0.1882	0.46	Declarative Question
25	105.44	0.0008	0.14	Statement
26	136.54	0.4257	0.45	Declarative Question
27	105.97	0.4241	0.21	Statement
28	84.87	0.0328	0.87	Declarative Question
29	10.14	2.662	1.72	Statement
30	122.54	0.293	0.57	Declarative Question

A_{aJ}: the amplitude of last accent command

A_{pI}: the magnitude of last phrase command

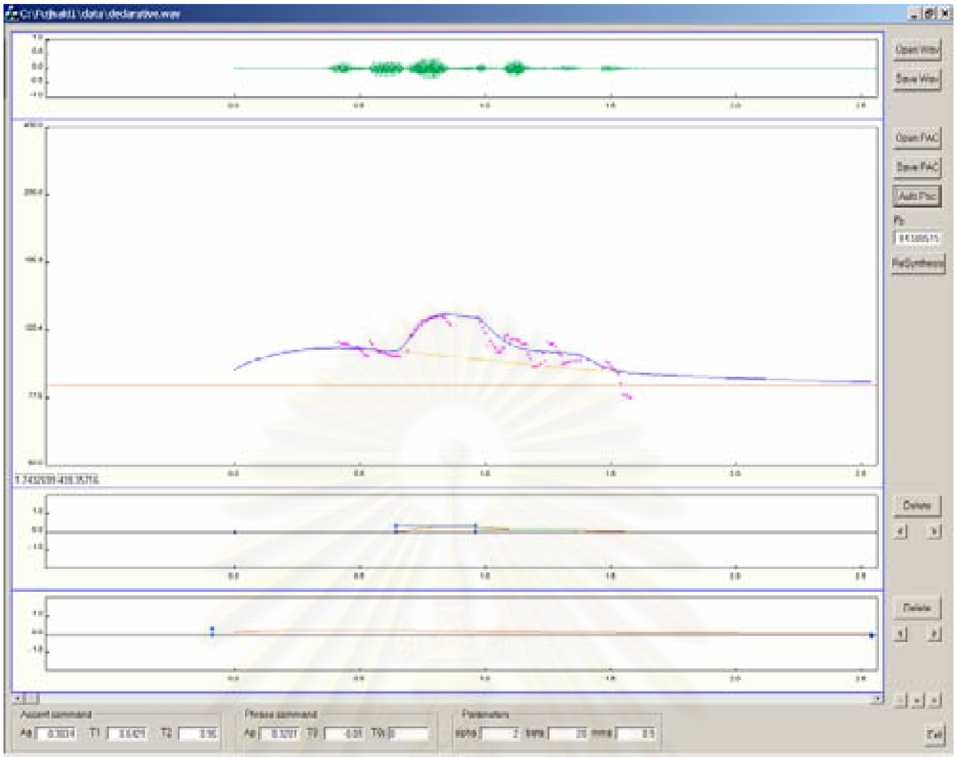
F_b: Baseline value of fundamental frequency

Table 3.2 Four combinations of Fujisaki-model parameters used to investigate the performance of the utterance-type recognizer

Combination	Fujisaki-model parameter(s)
I	A_{aJ}
II	A_{aJ} and $A_{p\mathrm{I}}$
III	F_b, A_{aJ} and A_{pI}
IV	F_b/100, A_{aJ} and A_{pI}

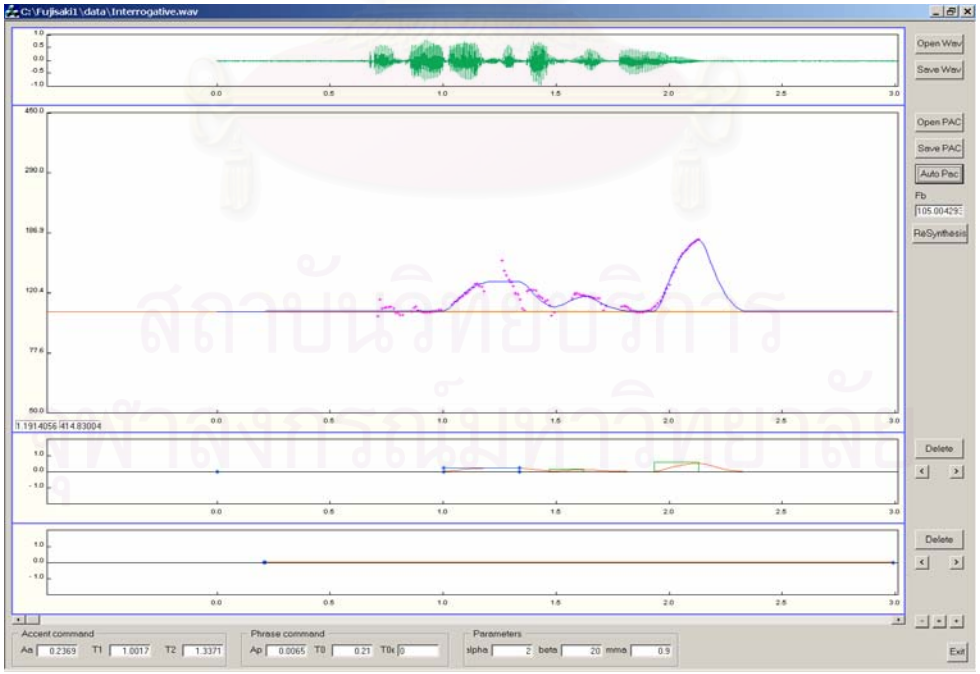
3.5 Experimental Results

Table 3.3 shows the recognition rates of the utterance-type recognizer using each combination of the Fujisaki-model parameters listed in Table 3-2 as the input of the neural networks. The recognition rate of the recognizer using only the amplitude of last accent command was 83.3%. The recognition rate of the recognizer increased to 90.0% when the recognizer used the combination of the amplitude of last accent command and the magnitude of last phrase command as the input of the neural network. When the recognizer used a combination of the baseline value of the fundamental frequency, the amplitude of last accent command and the magnitude of last phrase command as the input of the neural networks, the recognition rate of the recognizer decreased to 50.0%. The highest recognition rate was achieved when the recognizer used a combination of one percent of the baseline value of the fundamental frequency, the amplitude of last accent command and the magnitude of last phrase command as the input of neural network: 96.7%.



Accent command Aa 0.3034 T1 0.6421 T2 0.96
Phrase command Ap 0.3201 T0 -0.01 T0t 0
Parameters alpha 2 beta 20 mma 0.9

Figure 3.4 Waveform, pitch contour and Fujisaki-model parameters of an Indonesian Statement: “Dia sedang makan” (He is eating)



Fb 105.004293

Accent command Aa 0.2369 T1 1.0017 T2 1.3371
Phrase command Ap 0.0065 T0 0.21 T0t 0
Parameters alpha 2 beta 20 mma 0.9

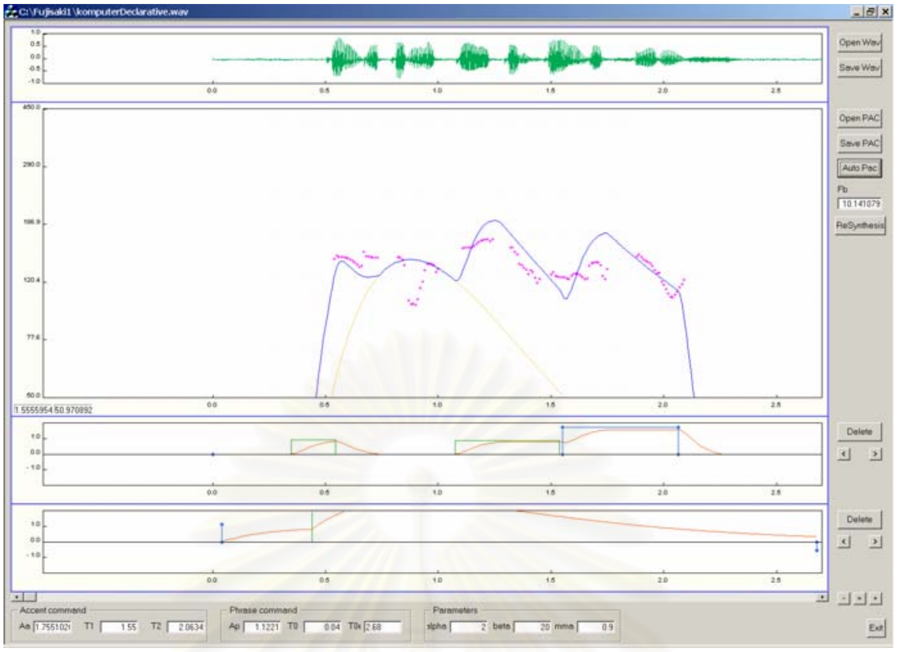
Figure 3.5 Waveform, pitch contour and Fujisaki-model parameters of an Indonesian declarative question: “Dia sedang makan?” (He is eating?)

Table 3.3 Recognition rate of the utterance-type recognizer based on Fujisaki model

Combination	Input of Neural Network	Recognition Rate (%)
I	A_{aJ}	83.3
II	A_{aJ} and $A_{p\mathrm{I}}$	90.0
III	F_b, A_{aJ} and A_{pI}	50.0
IV	$F_b/100$, A_{aJ} and A_{pI}	96.7

In this research, the neural network did not recognize all the patterns correctly because of the resemblance of the F0 contours of the declarative questions and the statements. Figure 3-4 and Figure 3-5 show the waveform, pitch contour, and Fujisaki-model parameters of a pair of a statement “Dia sedang makan” (He is eating) and its corresponding declarative question “Dia sedang makan?” (He is eating?). The figures show that the most important parts of the sentences to distinguish the two utterance-types are their final words. In the pair of the utterance-types, the declarative question has a final rise, while the statement has a final fall.

However, the pitch contours of the final words of another pair of declarative question and statement as shown in Figure 3.6 and Figure 3.7 are unlike the most patterns of those of the pairs of declarative questions and statements. Figure 3.6 and Figure 3.7 show another pair of an Indonesian statement “Komputer itu terjangkit virus” (The computer is infected by virus) and its corresponding declarative question “Komputer itu terjangkit virus?” (The computer is infected by virus?). In the pair, the last accent command of the statement is higher than that of its corresponding declarative question. It made the recognizer incorrectly distinguished the two utterance-types.



C:\Fujisaki1\komputerDeclarative.wav

Fb 10.141079

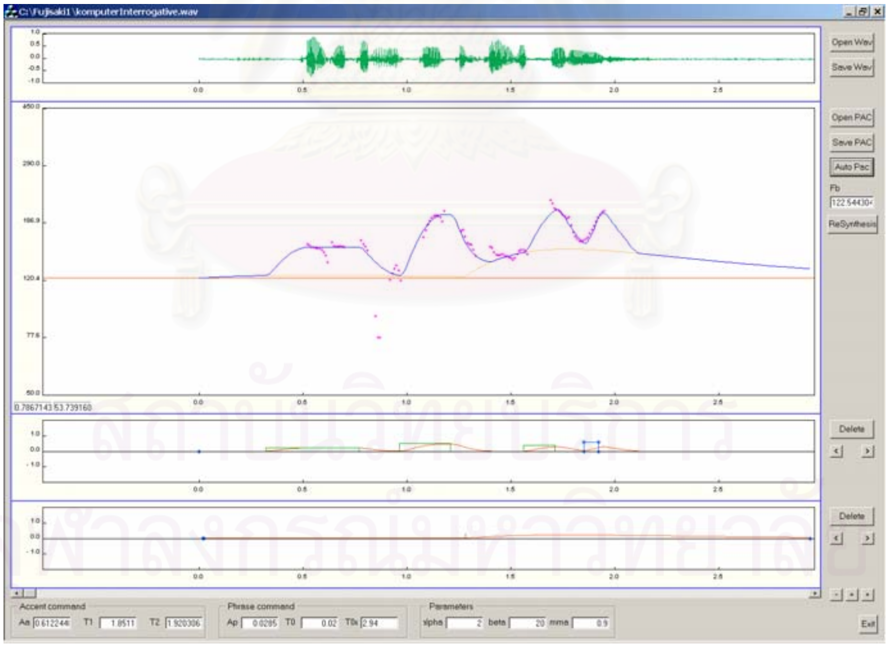
1.5555954 50.970892

Accent command: Aa 1.755102 T1 1.55 T2 2.0634

Phrase command: Ap 1.1221 T0 0.04 T0x 2.68

Parameters: alpha 2 beta 20 mma 0.9

Figure 3.6 Waveform, pitch contour and Fujisaki-model parameters of an Indonesian Statement: “Komputer itu terjangkit virus” (The computer is infected by virus)



C:\Fujisaki1\komputerInterrogative.wav

Fb 122.54430

0.7867143 53.739160

Accent command: Aa 0.612244 T1 1.8511 T2 1.920306

Phrase command: Ap 0.0285 T0 0.02 T0x 2.94

Parameters: alpha 2 beta 20 mma 0.9

Figure 3.7 Waveform, pitch contour and Fujisaki-model parameters of an Indonesian declarative question: “Komputer itu terjangkit virus?” (The computer is infected by virus?)

3.6 Summary

This chapter has discussed the design of an utterance-type recognizer based on Fujisaki model to distinguish the declarative questions and statements in Indonesian speech.

In this chapter, Fujisaki-model parameters were used as the features to classify the utterance type between the declarative questions and the statements in Indonesian speech. Four different combinations of Fujisaki-model parameters were used as the input of the classifier of the utterance-type recognizer. The highest recognition rate of the recognizer was achieved using a combination of a fraction value of F_b: F_b/100, the amplitude of last accent command, and the magnitude of last phrase command as the input of the neural networks.

CHAPTER IV
THE DESIGN OF AN ACOUSTIC MODEL OF AN INDONESIAN SPEECH RECOGNIZER

The hidden Markov model (HMM) as explained in chapter 2 has been known as a useful tool to create an acoustic model of a speech recognizer. There are many features to be the input of HMM. Among them are Mel Frequency Cepstral Coefficient (MFCC), linear frequency cepstrum coefficients (LFCC), linear prediction coefficients (LPC), reflection coefficients (RC), F0, energy, and duration (Davis and Mermelstein, 1980; Wang, 2001). MFCC has been widely used by many researchers as the features to create the acoustic model of the speech recognition (Maneenoi, 2004; Effendy et al., 2005). The MFCC combined with F0 have been used to represent the information of the source and the vocal tract respectively (Ezzaidi et al., 2004).

Several researchers already investigated the influence of the use of the features in the performance of a speech recognizer for other languages. Davis and Mermelstein (1980) described that the performance of the speech recognizer using MFCC as the features for the recognition is better than using LFCC, LPC or RC (Davis and Mermelstein, 1980). Matsumoto et al. (1998) showed that MFCC is slightly better than Mel-LPC in higher orders for female speakers, but slightly worse than Mel-LPC for male speakers (Matsumoto et al., 1998). Wang (2001) described that MFCC can improve the classification accuracy in addition to prosodic features in English stress classification. However, the gain using only prosodic features is greater than when only MFCC is used (Wang, 2001).

Besides of several investigations have been carried out in the influence of the features in a speech recognizer, however, none of them investigated the influence of the features in the Indonesian language. Many researchers of Indonesian speech recognizer used an acoustic model from other languages, such as English to be implemented in Indonesian speech recognizer (Martin et al., 2003; Wong et al., 2003; Sakti et al., 2004). In this research, an Indonesian acoustic model is designed and employed in an Indonesian speech recognizer. The investigation is limited on the features of the combination of MFCC with energy, their first order and second order derivatives. The acoustic model will be used in the automatic utterance segmentation module of an automatic utterance-type recognizer, which will be explained in chapter 6. The influence of the increment of the Gaussian mixtures in the performance of the acoustic model is investigated as well.

4.1 Mel Frequency Cepstral Coefficient (MFCC)

The spectral features can be obtained by passing the speech signal through a bank of bandpass filters. One of the main advantages of this approach is that the bandpass filters can be placed along the perceptual frequency scales such as critical band (Dautrich et al., 1983), bark scale (Ali et al., 2002), or Mel scale (Bu and Church, 2000). The filterbanks are generally triangular, and they are equally spaced along the Mel scale, which is defined as

$$Mel\,(f) = 2595\,\log_{10}(1 + \frac{f}{700}) \qquad (4.1)$$

Obviously, the Mel scale is linear below and logarithmic above 1 kHz. This scale is known to be a good scale for approximating the ability of human auditory system to discriminate frequencies.

To implement the filterbank, each segment of speech data is transformed using a Fourier transform and the magnitude is taken. Each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results are accumulated. If the cepstral parameters are computed from the log filterbank amplitude using the discrete cosine transform as shown in Equation (4.2), then, the Mel Frequency Ceptral Coefficients (MFCCs) are obtained.

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{i\pi}{N}(j - 0.5)\right) \qquad (4.2)$$

where, N is the number of filter bank channels and m_j is the log filterbank amplitude.

4.2 Speech Data

Speech data are recorded and collected from 20 Indonesian native speakers. They consist of 11 male and 9 female speakers with age ranging from 23 to 50. Each speaker was asked to utter twice 70 phone numbers and 89 sentences from Indonesian newspapers and Indonesian linguistics books, which consist of 1,305 words. Therefore, there are 280 utterances of phone numbers and 3,560 utterances of sentences, which consist of 52,200 words. The speech data are recorded with 16 kHz sampling rate and 16-bit resolution in an office environment.

4.3 Experimental Setup

Speech data were passed through a signal preprocessing routine consisting of signal pre-emphasis with a coefficient of 0.97 (Rabiner and Juang, 1993; Furui, 2001). A 25 msec Hamming window was applied every 10 msec in order to divide the speech signal into frames.

From the speech data, 49 sentences uttered by one male speaker and 1 sentence uttered by another male speaker was labeled manually based on phoneme. The Indonesian phonemes are listed in Table 4.1. The manual labeled speech data were used to create the acoustic model of the speech recognizer. A standard 5-state left-to-right Hidden Markov Model (HMM) with no skip state was employed in the design of the acoustic model. The manual acoustic model was used to automatically label the 335 phone numbers uttered by three male and two female speakers. The automatically labeled speech data were used to update the manual acoustic model to become an Indonesian acoustic model.

372 phone numbers uttered by two female speakers and four male speakers and 27 sentences uttered by one male speaker were used as the testing set to investigate the performance of the acoustic model of the Indonesian speech recognizer.

Table 4.1 Distinctive Feature Composition of Indonesian Phonemes (modified from Halim, 1981)

	i	u	o	a	e	ə	y	w	l	r	h	ʔ	p	b	f
Syllabic	+	+	+	+	+	+	-	-	+	+	-	-	-	-	-
Consonantal	-	-	-	-	-	-	-	-	+	+	-	-	+	+	+
Sonorant	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-
High	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-
Back	-	+	+	+	-	-	-	+	-	-	-	-	-	-	-
Low	-	-	-	+	-	-	-	-	-	-	+	+	-	-	-
Anterior	-	-	-	-	-	-	-	-	+	+	-	-	+	+	+
Coronal	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-
Round	-	+	+	-	-	-	-	+							
Tense	+	+	+	-	+	-	-	-							
Continuant									+	-	+	-	-	-	+
Voice									+	+	-	-	-	+	-
Nasal									-	-	-	-	-	-	-

	m	t	d	c	j	s	z	n	š	ɲ	k	g	x	ŋ	
Syllabic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
Consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
Sonorant	+	-	-	-	-	-	-	+	-	+	-	-	-	+	
High	-	-	-	-	-	-	-	-	+	+	+	+	+	+	
Back	-	-	-	-	-	-	-	-	-	-	+	+	+	+	
Low	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
Anterior	+	+	+	-	-	+	+	+	-	-	-	-	-	-	
Coronal	-	+	+	+	+	+	+	+	+	+	-	-	-	-	
Round															
Tense															
Continuant	-	-	-	-	-	+	+	-	+	-	-	-	+	-	
Voice	+	-	+	-	+	-	+	+	-	+	-	+	-	+	
Nasal	+	-	-	-	-	-	-	+	-	+	-	-	-	+	

Table 4.2 Six combinations of the acoustic-phonetic features used to create six types of Indonesian acoustic models

Combination	Symbol	Dimension of the acoustic-phonetic feature stream	Contents
I	MFCC	12	12 MFCCs
II	MFCC+E	13	12 MFCCs and Energy
III	MFCC+D	24	12 MFCCs and their first order derivatives
IV	MFCC+E+D	26	12 MFCCs, their energy and their first order derivatives
V	MFCC+D+A	36	12 MFCCs, their first and second order derivatives
VI	MFCC+E+D+A	39	12 MFCCs, their energy, their first and second order derivatives

Six types of the acoustic-phonetic features were used to create six types of acoustic models as shown in Table 4.2. The six types of Indonesian acoustic models were employed in an Indonesian speech recognizer. The performance of the Indonesian acoustic model was investigated from the recognition rate of the Indonesian speech recognizer. The number of the Gaussian mixtures per state of HMM was varied to investigate the best performance of the Indonesian acoustic model.

The performance of the acoustic model is computed using 2 formulas called "Percentage Correctness" and "Percentage accuracy" (Young et al., 2002). When the optimal alignment has been found, the number of substitution errors (S), deletion errors (D), and insertion errors (I) can be calculated. Then, the percentage correct is

$$\text{Percent Correct} = \frac{N-D-S}{N} \times 100\% \qquad (4.3)$$

where N is the total number of labels in the reference transcriptions. For many purposes, the percentage accuracy defined as

$$\text{Percent Accuracy } = \frac{N-D-S-I}{N}\times 100\% \qquad (4.4)$$

is a more representative figure of the recognizer performance.

4.4 Experimental Results

Figure 4.1 and Figure 4.2 show the percentage correct and the percentage accuracy of Indonesian speech recognizer. From the experimental results, the higher the number of the Gaussian mixtures, the higher both the accuracy and the percentage correct of the Indonesian speech recognizer. However, the increase of the number of Gaussian mixtures also increased the computational complexity. The Indonesian speech recognizer with 16 Gaussian mixtures achieved the highest percentage accuracy and the highest percentage correct.

From the six combinations of the MFCCs with energy, their first and second order derivatives, the Indonesian acoustic model of the speech recognizer using only MFCCs achieved the lowest both of the percentage correct and the percentage accuracy. However, the Indonesian speech recognizer with only MFCCs had the smallest computational complexity. The addition of the energy to MFCC increased the percentage correct of the Indonesian speech recognizer to be 72.4%, 77.3%, 81.6%, 84.9% and 86.4% for one, two, four, eight and sixteen Gaussian mixtures, respectively. The addition of the first order derivatives of MFCCs to MFCCs increased the percentage correct of the Indonesian speech recognizer higher than the addition of the energy to MFCCs did, i.e. to be 84.8%, 88.3%, 89.60%, 90.1%, and 90.2% for one, two, four, eight and sixteen Gaussian mixtures, respectively. A combination of MFCCs with energy and their first order derivatives increased the percentage correct of the Indonesian speech recognizer higher than the combination of MFCC and their first order derivatives did, i.e. to be 85.0%, 88.4%, 89.7%, 90.4%, and 90.6% for the one, two, four, eight and sixteen Gaussian mixtures, respectively. The combination of MFCCs with energy, their first and second order derivatives increased the percentage correct of the Indonesian speech recognizer higher than the combination of MFCCs with energy and their first order derivatives did for four, eight and sixteen Gaussian mixtures, i.e. 88.8%, 90.6%, and 91.3%, respectively. The combination of MFCCs with their first and second order derivatives made the percentage correct of the Indonesian speech recognizer to be 83.4% and 86.9% for one and two Gaussian mixtures, respectively.

The effect of the combinations of the MFCCs with energy, their first and second order derivatives on the percentage accuracy of the speech recognizer is the same with their effect on the percentage correct. The addition of energy to MFCCs increased the percentage accuracy of the Indonesian speech recognizer to be 63.6%, 70.6%, 76.8%, 81.4%, and 83.8% for one, two, four, eight and sixteen Gaussian mixtures, respectively. The addition of the delta to MFCCs increased the percentage accuracy of the Indonesian speech recognizer higher than the combination of MFCCs with energy did, i.e. to be 80.2%, 84.2%, 86.0%, 87.1%, and 87.6% for one, two, four, eight and sixteen Gaussian mixtures, respectively. The combination of MFCCs with energy and their first order derivatives increased the percentage accuracy of the Indonesian speech recognizer higher than the combination of MFCCs with their first order derivatives did, i.e. to be 81.2%, 85.2%, 86.8%, 87.8 and 88.3% for one, two, four, eight and sixteen Gaussian mixtures, respectively. A combination of MFCCs with their first and second order derivatives increased the percentage accuracy higher than the combination of MFCCs with energy, and their first order derivatives did for the acoustic model with 16 Gaussian mixtures, i.e. to be 88.7%. A combination of MFCCs with energy, their first and second order derivatives increased the percentage accuracy of the Indonesian speech recognizer higher than the combination of MFCCs with energy, and their first order derivatives did for four, eight and sixteen Gaussian mixtures, i.e. to be 87.1%, 89.3%, and 89.7%, respectively. The combination of MFCCs with energy, their first and second order derivatives increased the percentage correct of the Indonesian speech recognizer to be 79.6% and 83.7% for one and two Gaussian mixtures, respectively.

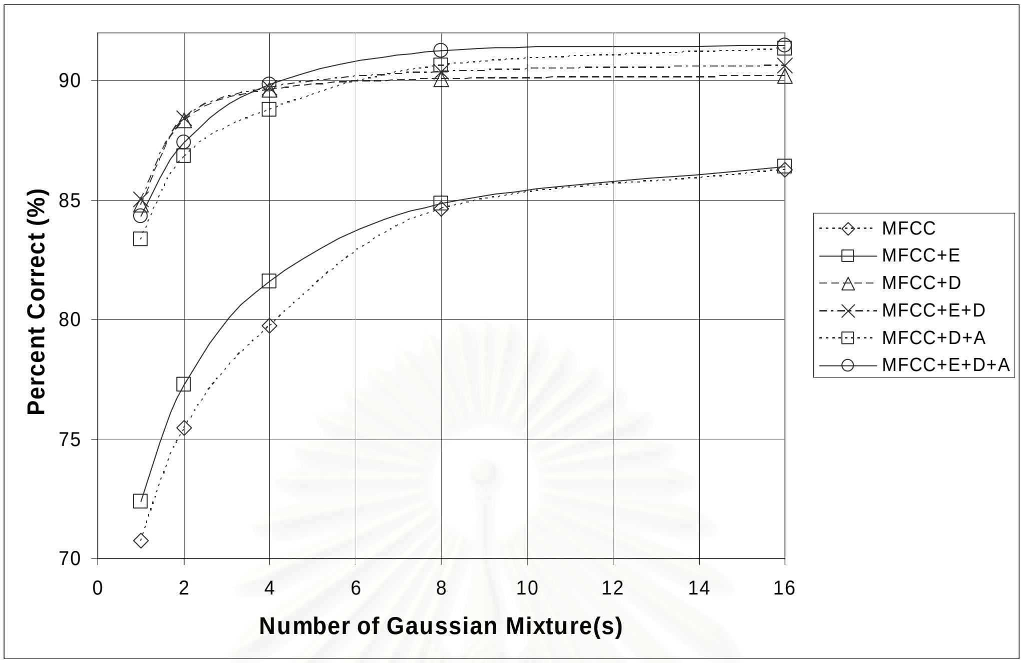


Figure 4.1 Percentage correct of Indonesian speech recognizer.

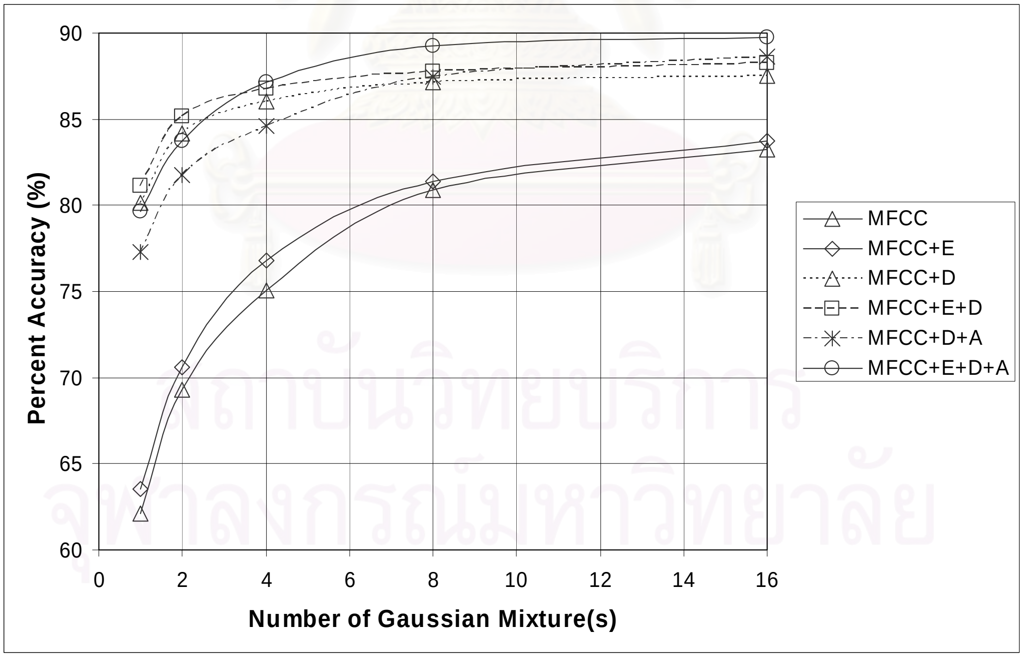


Figure 4.2 Percentage accuracy of Indonesian speech recognizer.

The highest of both the percentage correct and the percentage accuracy of the speech recognizer with either one or two Gaussian mixtures were achieved using the combination of MFCCs with energy, and their first order derivatives. The highest both the percentage correct and the percentage accuracy of the speech recognizer using four, eight or 16 Gaussian mixtures were achieved using the combination of MFCC with energy, their first and second order derivatives. In the research, the highest percentage correct and the highest percentage accuracy of the speech recognizer were achieved using 16 Gaussian mixtures and the combination of MFCCs with energy, their first and second order derivatives.

4.5 Summary

This chapter has discussed the study on the design of an acoustic model of Indonesian speech recognizer. The Indonesian acoustic model yielded from the research will be employed in the design of an automatic utterance-type recognizer to distinguish declarative question and statement in Indonesian speech, which will be further described in Chapter 6.

In this research, the higher the number of Gaussian mixtures used in the design of the Indonesian acoustic model, the higher both the percentage correct and the percentage accuracy of the speech recognizer. The highest percentage correct and the highest percentage accuracy of the speech recognizer using either one or two Gaussian mixtures were achieved using the combination of MFCCs with energy and their first order derivatives. The highest percentage correct and the highest percentage accuracy of the speech recognizer using four, eight or sixteen Gaussian mixtures were achieved using the combination of MFCCs with energy, their first and second order derivatives. In the research, the highest percentage correct and the highest percentage accuracy of the speech recognizer were achieved using 16 Gaussian mixtures and the combination of MFCCs with energy, their first and second order derivatives.

CHAPTER V
FINAL WORD OF DECLARATIVE QUESTION AND STATEMENT IN INDONESIAN SPEECH

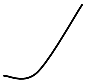
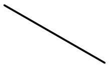
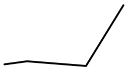
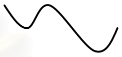
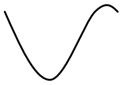
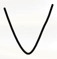
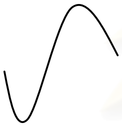
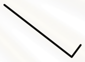
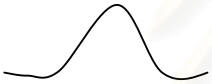
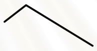
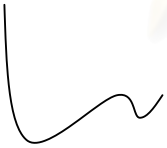
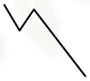
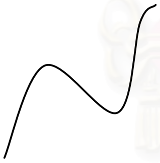
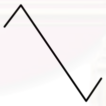
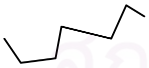
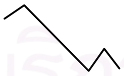
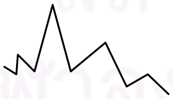
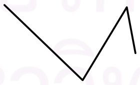
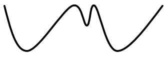
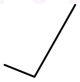
The characteristics of the final word of declarative question and statement in Indonesian speech were an open question. Ebing (1997) explained the form and the function of pitch movements in Indonesian (Ebing, 1997) but not including the characteristics of the final word of an utterance types in Indonesian speech. The characteristics of the final word of declarative question and statement in Indonesian speech had not been investigated. I have a hypothesis that the final word of declarative question and statement in Indonesian speech might be used to distinguish the two utterance types. To confirm the hypothesis, the characteristics of the final words of each utterance type will be investigated.

This chapter describes the investigation of the characteristics of the final words of declarative questions and statements in Indonesian speech. The investigation includes the pitch contour, pitch range, maximum pitch and speech rate of the sentence's final word.

5.1 Speech Data

Speech data are recorded from 35 Indonesian native speakers. They consist of 11 female and 24 male speakers with age ranging from 23 to 50. The subjects are asked to utter 29 pairs of sentences as naturally as possible. The pair of sentences consists of statements and their corresponding declarative questions. The sentences are selected from the daily life conversation among Indonesian speakers (Effendy et al., 2004). From the recording, there are 2030 utterances in the speech data. The speech data are recorded in an office environment. The recorded speech data are digitized at 16 kHz sampling rate and 16-bit resolution.

Table 5.1 Types of pitch contour of the final word extracted from 290 pairs of declarative questions and statements in Indonesian Speech.

No	Declarative Question	Percentage of occurrence	Statement	Percentage of occurrence
1		5.86		6.21
2		8.62		31.03
3		14.48		1.03
4		8.28		1.38
5		5.52		13.10
6		2.07		26.89
7		33.45		5.17
8		5.17		8.62
9		3.10		6.21
10		13.45		0.34

5.2 Pitch of the Final Words

Pitch is the major correlate of intonation (Yan et al., 2003; Watson and Hughes, 2006) and has been used by many researchers in prosody study. In this section, the pitch contour, pitch range and maximum pitch are investigated on their relation with the utterance type of declarative question and statement of Indonesian speech.

5.2.1 Experimental Setup

The pitch contour of the sentences were extracted using PRAAT software (Boersma and Weenink, 2004) to produce the pitch data. The pitch contour of the final words was produced using both the pitch contour extracted using PRAAT software and final word boundary information extracted with the manually segmentation of the utterances. The manual utterance segmentation to determine the final word boundary was performed to avoid the error in the segmentation.

To get the types of the pitch contours of the final words of Indonesian sentences, 290 declarative questions and 290 statements from the speech data were analyzed. Similar forms of the pitch contours were grouped into the same type. The pitch range and maximum pitch of the sentence final words were calculated automatically from the pitch contour.

5.2.2 Experimental Results

5.2.2.1 Pitch Contour

Table 5.1 shows the types of the pitch contour of the sentence's final word. From the experiments, it was found that there were ten types of the pitch contour of the final word of both declarative questions and statements. 33.45% of the pitch contours of the final word of the declarative question have a 'rising wave' form. 14.48% of the pitch contours have a 'falling-rising' form. 13.45% of the pitch contours have a 'falling-rising-falling-rising' form. Others have other forms such as 'rising' (5.86%), 'end rising' (8.62%), 'falling-slightly rising' (2.07%). In contrast with declarative questions, 31.03% of the pitch contours of the final word of the statements have a ‘falling wave’ form. In general, 92.4% of the pitch contours of the final word of the statements have a positive general slope, whereas only 61.38% of the declarative questions have a negative general slope.

Figure 5.1 shows the pitch contours of the three-syllable-final-word of Indonesian declarative questions from three sentences each contains either the final words ‘pen-ja-hat’ (‘criminal’), ‘sen-di-ri’ (‘alone’), or ‘me-na-ri’ (‘dance’). Part (a) and (b) of Figure 5.1 display the pitch contours of the final word having stress located at the final syllable (‘syl3’) uttered by male and female speakers respectively. The positions of syllables ‘syl1’, ‘syl2’, and ‘syl3’ in Figure 5.1 are not accurate for each speaker. They show only the common position, because each speaker can utter the sentence with different speech rate and different duration for each syllable.

Like Figure 5.1, Figure 5.2 also shows the pitch contours of the three-syllable-final-word of Indonesian declarative questions from three sentences each contains either final words ‘pen-ja-hat’ (‘criminal’), ‘sen-di-ri’ (‘alone’), or ‘me-na-ri’ (‘dance’) but the final words have stress located at the second syllable (‘syl2’) of the words. Figure 5.1 and Figure 5.2 confirm that in Indonesian speech, the final words of the declarative questions do not have only one form of the pitch contour. Different position of the stressed syllable changes the form of the pitch contours.

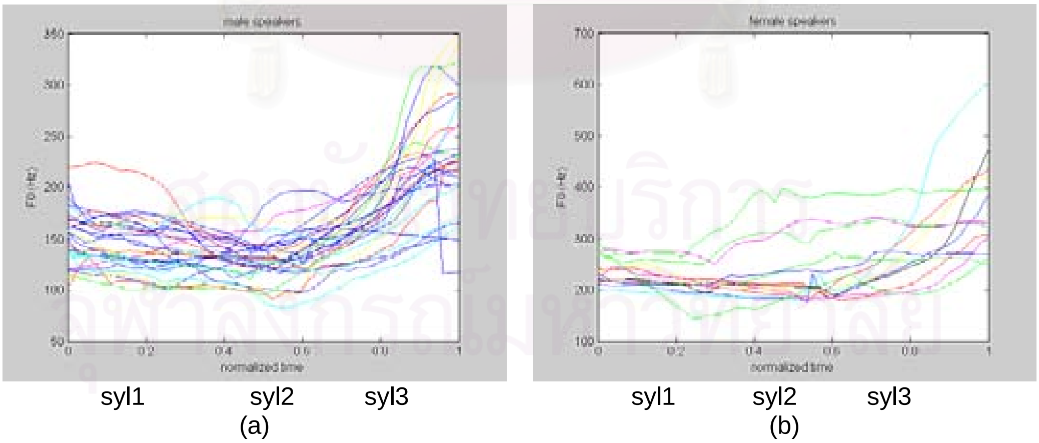


Figure 5.1 Pitch contours of the three-syllable-final-word of declarative questions uttered by (a) male and (b) female speakers with the stress on the last syllable. Each line represents the pitch contour of each speaker.

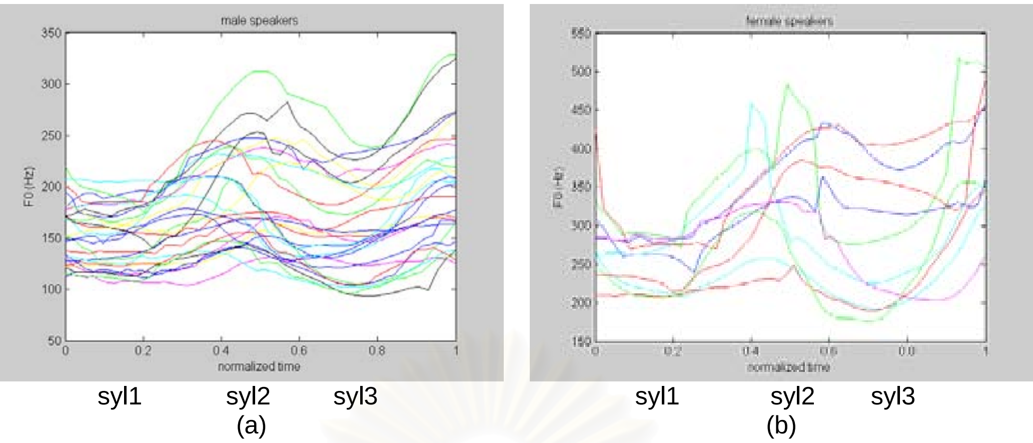


Figure 5.2 Pitch contours of the three-syllable-final-word of declarative questions uttered by (a) male and (b) female speakers with the stress on the second syllable. Each line represents the pitch contour of each speaker.

Figure 5.3 depicts the pitch contours of the two-syllable-final-word of declarative questions with different stressed syllables in a normalized time domain. Part (a) and (b) of Figure 5.3 display the pitch contour of the final word ‘syl1-syl2’ with stress placed on the last-syllable for male and female speakers respectively. Part (c) and (d) of Figure 5.3 show the pitch contour of the final word ‘syl1-syl2’ having the pitch contour as ‘rise-fall-rise’. The label ‘syl1’ and ‘syl2’ in the figures are placed in a common location for each speaker, because each speaker uttered the sentence naturally without using the same speech rate for each syllable. The stress on different syllable of the final word produces different contours of the pitch.

The examples of the pitch contours of the two-syllable-final-word and the three-syllable-final-word of statements uttered by male and female speakers are separately displayed in Figure 5.4. Most of the pitch contours have a falling form. The rest have a rising form.

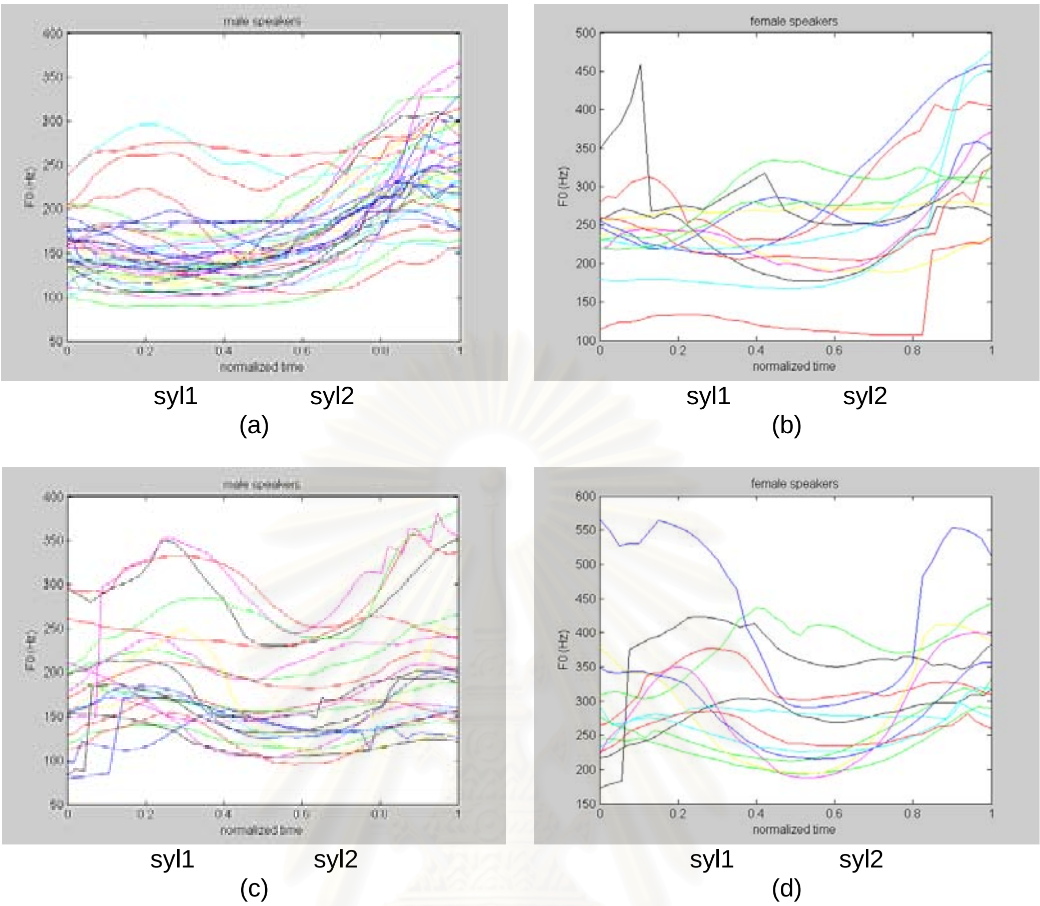
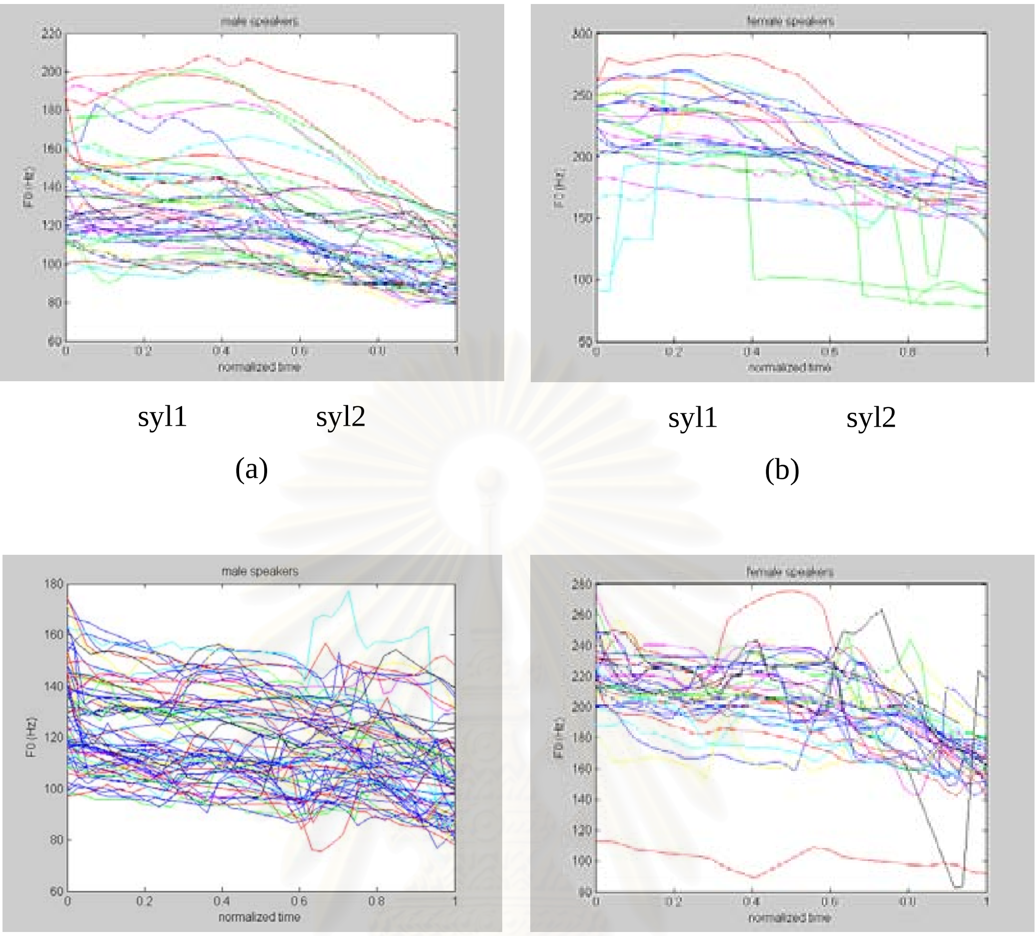


Figure 5.3 Pitch contours of the two-syllable-final-word of declarative questions. Different lines are used for different speakers.

Figure 5.1, Figure 5.2, Figure 5.3, and Figure 5.4 confirm that both the declarative questions and the statements uttered by female speakers have higher average pitch than those uttered by male speakers. The findings are the same with the findings of other researchers for other languages. This characteristic may be universal for many languages. The pitch contours of the final word of each utterance type have various forms. This characteristics increase the difficulty in the utterance type recognition using the pitch contour of sentence’s final word.



syl1 syl2

(a)

syl1 syl2

(b)

syl1 syl2 syl3

(c)

syl1 syl2 syl3

(d)

Figure 5.4 Pitch contours of the two-syllable-final-word (top) and the three-syllable-final-word (below) of statements. Different lines are used for different speakers.

5.2.2.2 Pitch Range

Figure 5.5 and Figure 5.6 show the histogram of the pitch range of the final word of declarative questions and statements uttered by male and female speakers respectively. Although most of the pitch ranges of statements are smaller than those of declarative questions, there is intersection between them that makes error in the recognition of the two utterance types when the utterance-type recognizer uses only the pitch range information to distinguish the two utterance types.

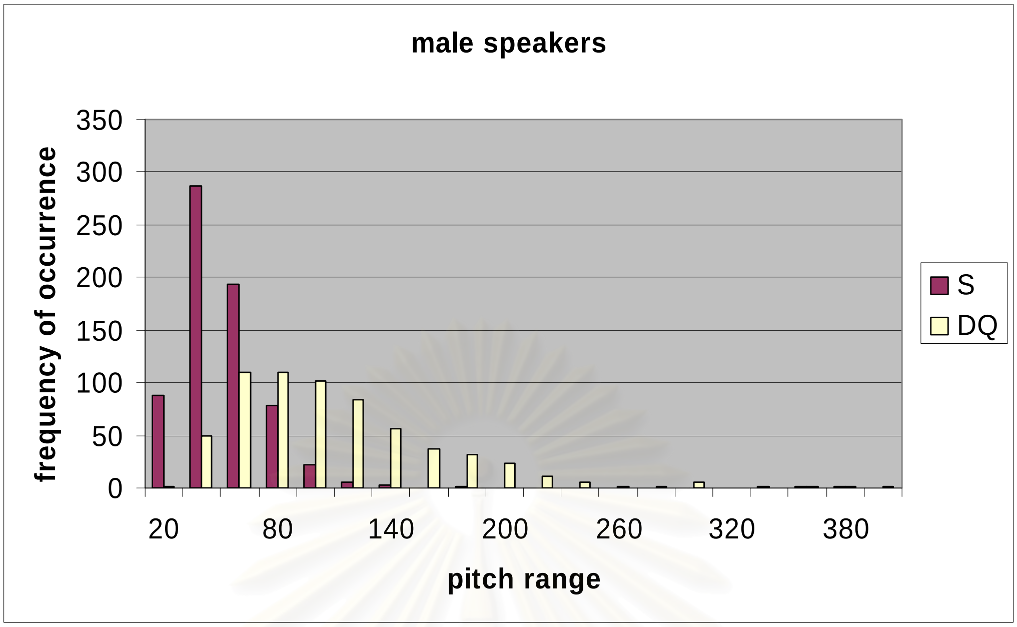


Figure 5.5 Histogram of the pitch range of the final word of declarative questions and statements uttered by male speakers

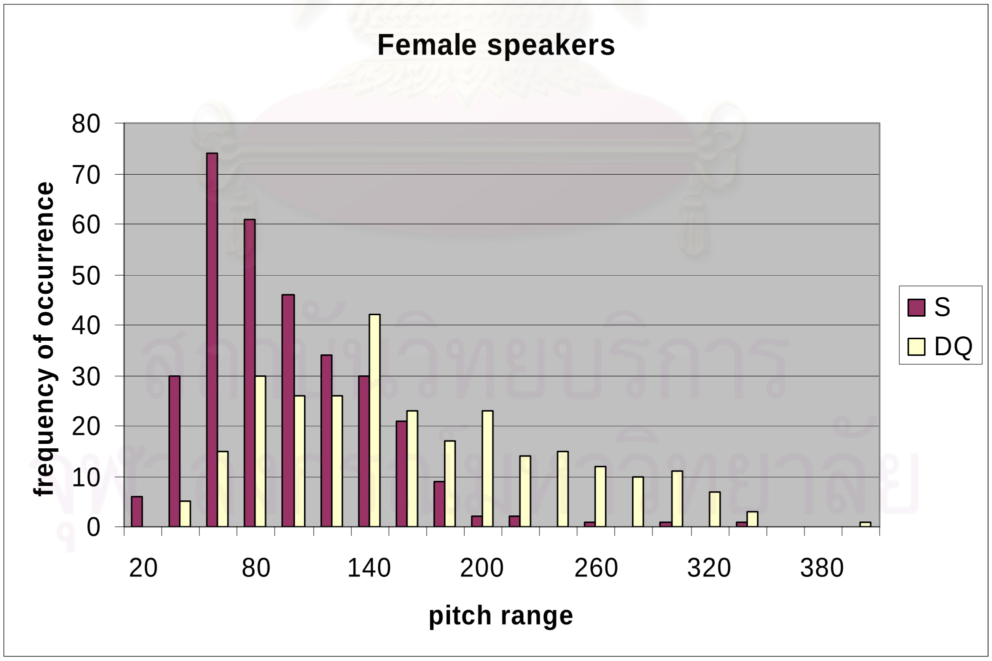


Figure 5.6 Histogram of the pitch range of declarative questions and statements uttered by female speakers

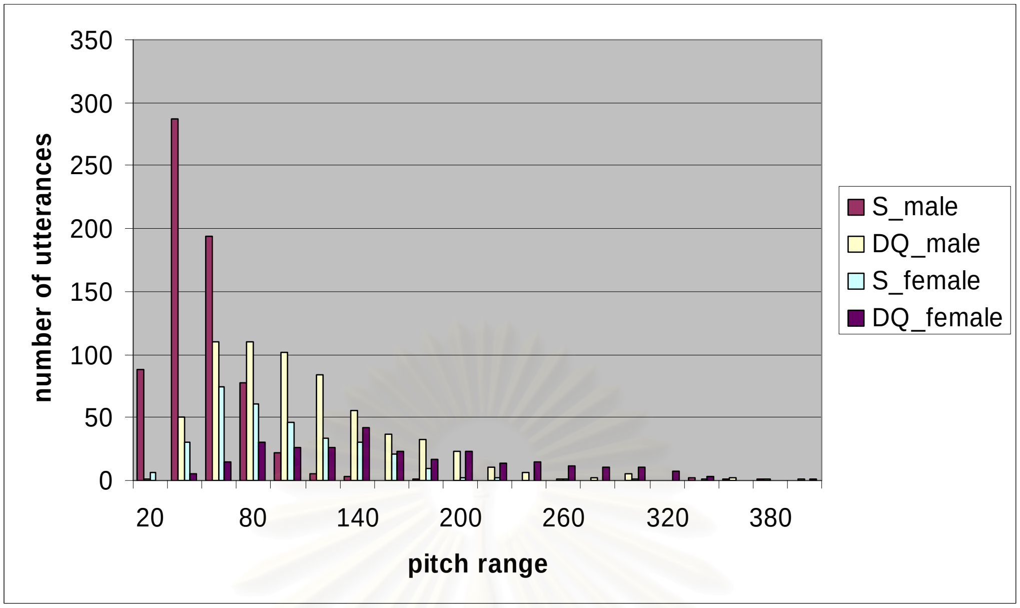


Figure 5.7 Histogram of the pitch range of declarative questions and statements uttered by male and female speakers

Moreover, there is a large intersection between the pitch range of the final words of the declarative questions of male speaker and the pitch range of the final words of the statements of female speakers as shown in Figure 5.7. Consequently, there will be error in the gender-dependent utterance-type recognition when the recognizer uses only the pitch range information to distinguish the two utterance types.

Table 5.2 The average and standard deviation of the pitch range of the final word of declarative questions and statements uttered by male and female speakers

pitch range	male speaker		female speaker	
	statement	declarative-question	statement	declarative-question
average	42.72	99.59	85.56	154.54
standard deviation	31.28	54.57	44.90	74.68

Table 5.2 shows the average and standard deviation of the pitch range of the final word of declarative questions and statements uttered by male and female speakers. Female speakers tend utter declarative questions with a larger average pitch range than male speakers do. Both male and female speakers utter declarative questions with a larger pitch range than statements.

5.2.2.3 Maximum Pitch

Figure 5.8, Figure 5.9, and Figure 5.10 show the histogram of the maximum pitch of declarative questions and statements uttered by male, female, and both male and female speakers respectively. Female speakers tend to utter both declarative questions and statements with higher maximum pitch than male speakers do. However, the distribution of the maximum pitch of the final words of declarative questions uttered by male speakers and the maximum pitch of the final words of the statements uttered by female speakers has a large intersection.

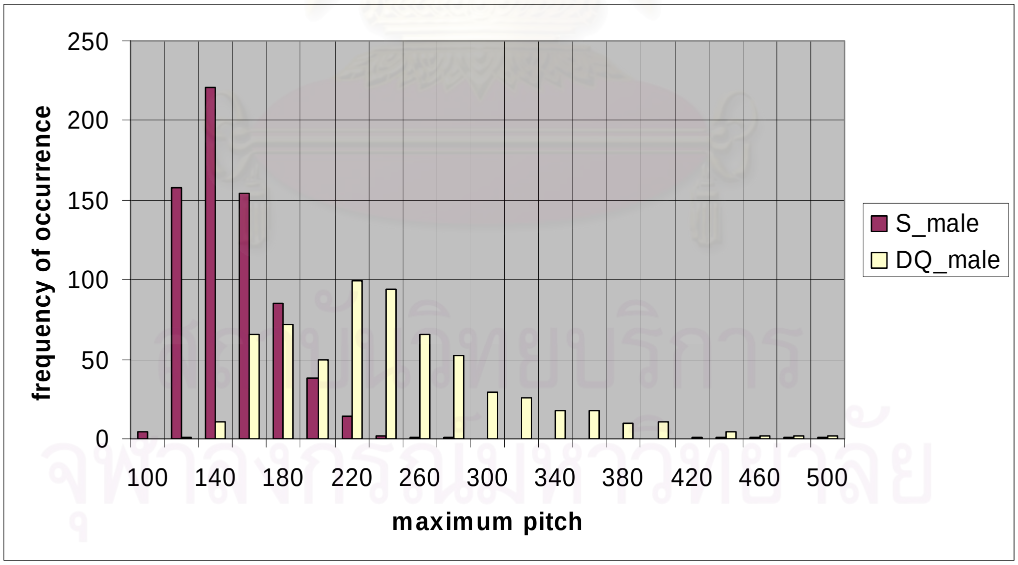


Figure 5.8 Histogram of the maximum pitch of declarative questions and statements uttered by male speakers

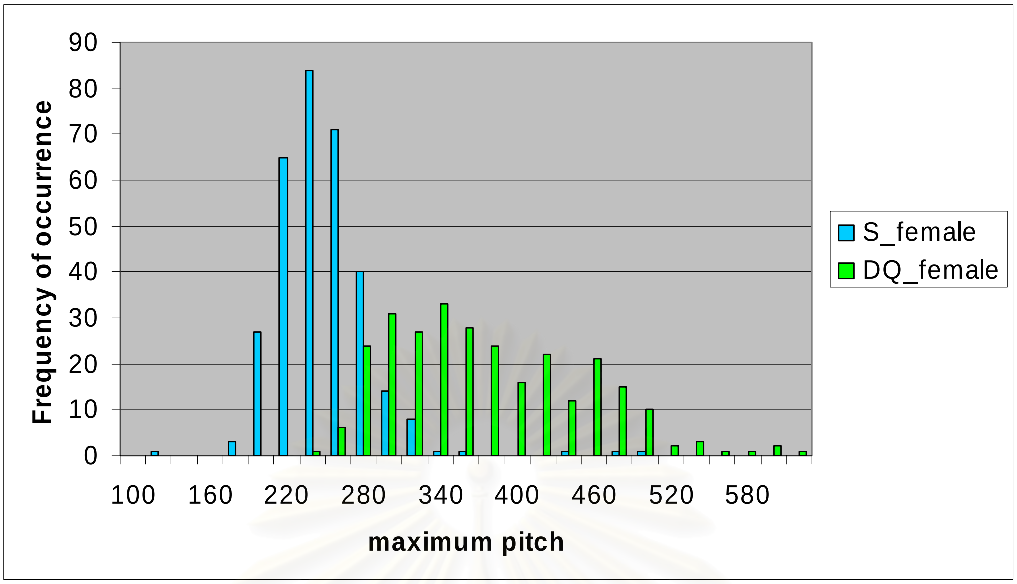


Figure 5.9 Histogram of the maximum pitch of declarative questions and statements uttered by female speakers

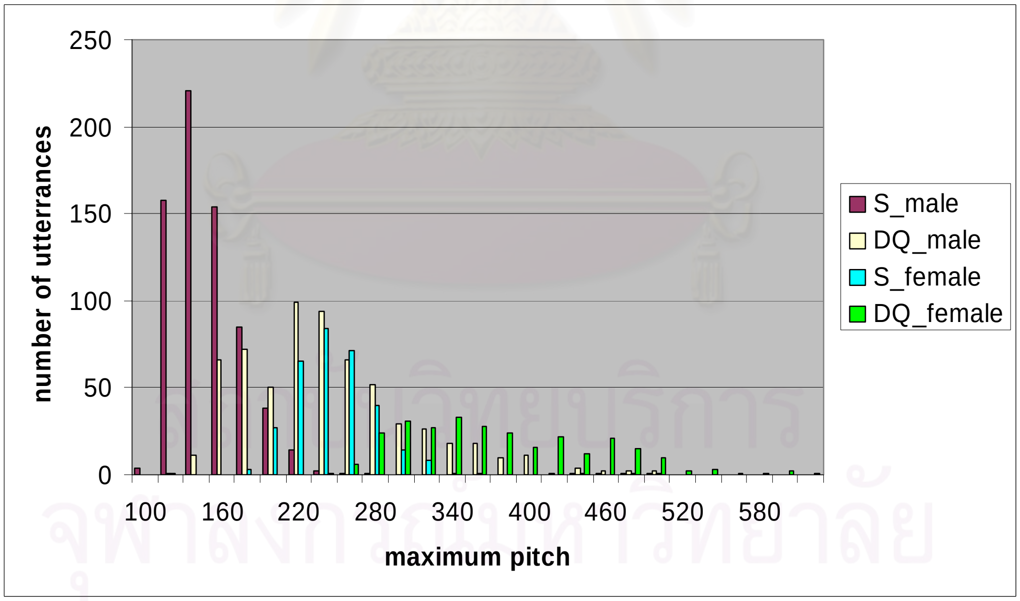


Figure 5.10 Histogram of the maximum pitch of declarative questions and statements uttered by male and female speakers

Table 5.3 The average and standard deviation of the maximum pitch of the final words of the declarative questions and the statements uttered by male and female speakers

maximum pitch	male speaker		female speaker	
	statement	declarative-question	statement	declarative-question
average	142.43	232.57	238.43	366.61
standard deviation	35.89	65.59	37.04	74.46

Table 5.3 shows the average and standard deviation of the maximum pitch of the final words of declarative questions and statements uttered by male and female speakers. The averages of the maximum pitch values of male-statements and female-declarative-questions are close, i.e. 232.57 Hz and 238.43 Hz respectively. It makes a gender independent utterance-type recognizer will misrecognizes some of the declarative questions and statements when the recognizer uses only the maximum pitch information to distinguish the two utterance types.

5.3 Speech Rate

5.3.1 Experimental Setup

The sentence final words were segmented manually. The speech rate of each final word was measured from its duration and the number of syllables in the final words.

5.3.2 Experimental Results

Figure 5.11 shows the speech rates of the final word of 18 Indonesian declarative questions and those of the statements uttered by four Indonesian female speakers.

The figure suggests that most of the final words of declarative questions in Indonesian speech are spoken at a slower rate their corresponding statements.

Figure 5.12 shows the histogram of the duration per syllable of the final word of Indonesian declarative questions and statements. From the figure, the duration per syllable of the final word of the statements and the declarative questions have a similar distribution. Therefore, the duration per syllable may be a low correlate to distinguish Indonesian statements and declarative questions.

5.4 Summary

This chapter has described the investigation of the characteristics of the final word of declarative questions and statements in Indonesian speech. They include the pitch contour, pitch range, maximum pitch and the speech rate of the final words of the declarative questions and statements.

Both the pitch range and maximum pitch of the final words of declarative questions are larger than those of the statements are. However, there is intersection of the distribution of the parameters. Consequently, an utterance-type recognizer will misrecognize some of the declarative questions and the statements when it uses only either the pitch range or the maximum pitch information.

Most of the final words of declarative questions are spoken at a slower rate than the final words of their corresponding statements. The duration per syllable of the final word of the statements and the declarative questions has a similar distribution. Therefore, the duration per syllable may be a low correlate to distinguish Indonesian statements and declarative questions.

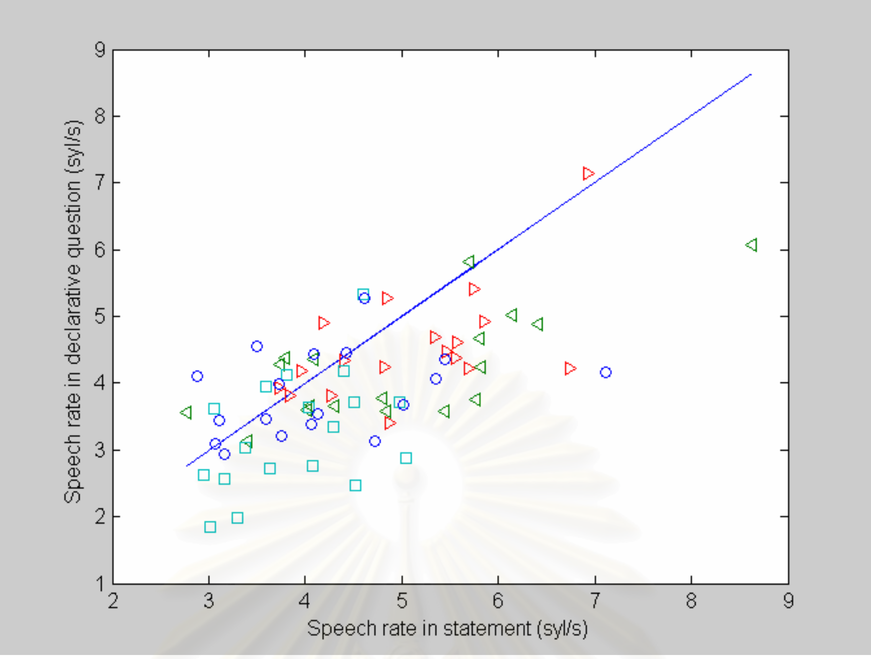


Figure 5.11 Relation of the speech rates (syl/s) of the final word of 18 pairs of Indonesian statements and their corresponding declarative questions uttered by four female speakers. Each symbol of the points is for each speaker.

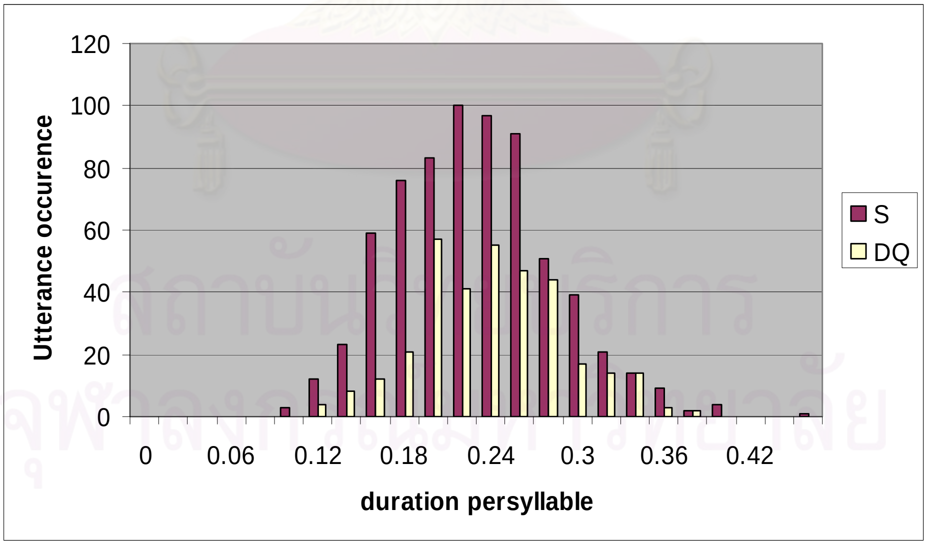


Figure 5.12 Duration per syllable of the final words of Indonesian Statement (S) and Declarative Question (DQ).

CHAPTER VI
AUTOMATIC UTTERANCE-TYPE RECOGNIZER USING THE POLYNOMIAL COEFFICIENTS OF THE PITCH CONTOUR OF SENTENCE'S FINAL WORD

In Chapter 3, the parameters of Fujisaki model are used to represent declarative questions and statements in Indonesian speech (Effendy et al., 2004). The utterance-type recognizer was designed as speaker-dependent system. In this chapter, an automatic utterance-type recognizer, which is a speaker-independent system and covers larger speech data, is proposed. The proposed utterance-type recognizer uses polynomial coefficients to represent the pitch contour of the sentence final words because the algorithm to estimate the parameters of Fujisaki model used in Chapter 3 in representing the pitch contour is too complicated to be implemented in an automatic utterance-type recognizer.

Prior to modeling a given F0 contour using Fujisaki model, two tasks are performed: (1) Intermediate F0 values for unvoiced speech segments and short pauses are interpolated from the extracted F0 contour, (2) Microprosodic variations caused by the influence of individual speech sounds are smoothed out, as the Fujisaki model explicitly deals with macroprosody only (Mixdorff et al., 2003). The Fujisaki model produces a particular F0 contour in the log F0 domain by superimposing three components: the phrase component which corresponds to the phrase-wise slow overall declination line, the accent component made up by the faster movements in the F0 contour connected with syllabic accents, and Fb, a speaker-individual constant. In order to separate the accent component from the phrase component and Fb, the smooth contour is passed through a high pass filter with a stop frequency at 0.5 Hz. The output of the high-pass (henceforth called 'high frequency contour' or HFC) is subtracted from the smooth contour yielding a 'low frequency contour' (LFC), containing the sum of phrase component and Fb.

In a sequence of phrase commands, the onset of a new command is characterized by a local minimum in the phrase component. Consequently, the LFC is searched for local minima, applying a minimum distance threshold of 1 sec between consecutive phrase commands. For initializing, the magnitude value Ap assigned to each phrase command part of the LFC after the potential onset time T0 of a phrase command is searched for the next local maximum. Ap is then calculated in proportion to the frequency value found at this point. The time constant α is set to 2.0/sec, a value found appropriate after a series of preliminary trials. In order to yield an optimal initialization of phrase commands, the F0 contour is examined for pauses longer than 400 msec. Since the minima of the LFC do not provide the exact locations of upcoming phrase commands, phrase commands found in the vicinity of a pause are readjusted and aligned with the pause. For initializing the appropriate number, polarity and onset times T1 and offset times T2 of accent commands, the cubically smoothed contour pertaining to the part of the wave file between first and last voiced frame, is subdivided into segments with positive or negative gradient, respectively. These contour segments are searched for points where the derivative exhibits a maximum, that is, the inflection points of the cubically smoothed curve. Inflection points on contour segments with rising slope are associated with the offset of a negative accent command and the subsequent onset of a positive accent command, and inflection points on contour segments with falling slope are associated with the offset of a positive accent command and the subsequent onset of a negative accent command. Hence, an alternating sequence of positive and negative commands is yielded initially. The HFC is basically DC-free and therefore oscillates around 0. For initializing the accent command amplitude Aa, the positive or negative maximum in the HFC between the initial settings of T1 and T2 is determined and Aa set in proportion to the frequency value found at this point. Accent commands are not continued across major pauses in the speech signal. The accent command time constant β is set to a value of 20/sec.

The automatic utterance-type recognizer proposed in this chapter uses the polynomial expansion to extract the polynomial coefficients from the pitch contour of the sentence final words. The algorithm to find the polynomial coefficients has a smaller computational complexity than the algorithm to find the parameters of Fujisaki model as will be explained in section 6.2.4. The polynomial coefficients are used as the input of a classifier to distinguish the declarative questions and the statements. The automatic utterance-type recognizer is optimized using various numbers of hidden nodes in the neural networks and various orders of polynomial expansion.

6.1 Speech Data

The speech data were 29 sentence pairs, each comprising a statement and its corresponding declarative question. The two sentences in each pair have the same words in the same order; they differ only in intonation. This is different from the study of Yuan et al. (Yuan et al., 2002), where the number of words in each sentence differed from each other. The sentences are selected from the daily-life conversations among Indonesian speakers (Effendy et al., 2004) as listed in Appendix A. Speech data of 35 Indonesian native speakers are recorded. They consist of 11 female and 24 male speakers with ages ranging from 23 to 50. The speech data are recorded in an office environment. In the recording, each subject was asked to utter the 29 pairs of sentences as naturally as possible. After the recording session, the speech data are verified perceptively three times. Utterances with wrong intonation are removed.

Then, the speech data were divided into four sets with the balance in gender and data amount maintained as illustrated in Table 6.1. Sets I and IV contain speech data from 12 male speakers and 5 female speakers. Sets II and III contain speech data from another group of speakers; 12 male speakers and 6 female speakers. Sets I and II contain speech data consisting of 14 pairs of sentences, while sets III and IV contain speech data consisting of 15 pairs of sentences from another group. Consequently, there are 1866 utterances consisting of 221 statements and 221 declarative questions in set I, 234 statements and 234 declarative questions in set II, 241 statements and 241 declarative questions in set III, and 237 statements and 237 declarative questions in set IV.

Table 6.1 The Indonesian speech database of statements and declarative questions

		Speakers	
		First group - 12 males - 5 females	Second group - 12 males - 6 females
Sentences	14 pairs of sentences in the first group	Set 1 statements: 216 declarative questions: 216	Set 2 statements: 230 declarative questions: 230
	15 pairs of sentences in the second group	Set 4 statements: 228 declarative questions: 228	Set 3 statements: 240 declarative questions: 240

6.2 Automatic Utterance-type recognizer

An automatic utterance-type recognizer is proposed to distinguish statements and declarative questions in Indonesian speech. This recognizer is speaker- and gender-independent, and consists of an automatic utterance segmentation module, an *F0* extractor, a normalizer, a feature extractor, and a classifier as illustrated in Figure 6.1. In this study, it is assumed that the correct transcription is given, since there are no large databases available to train accurate acoustic models in the Indonesian language at present. Each of the subsystem of the recognizer will be described further in the next subsection.

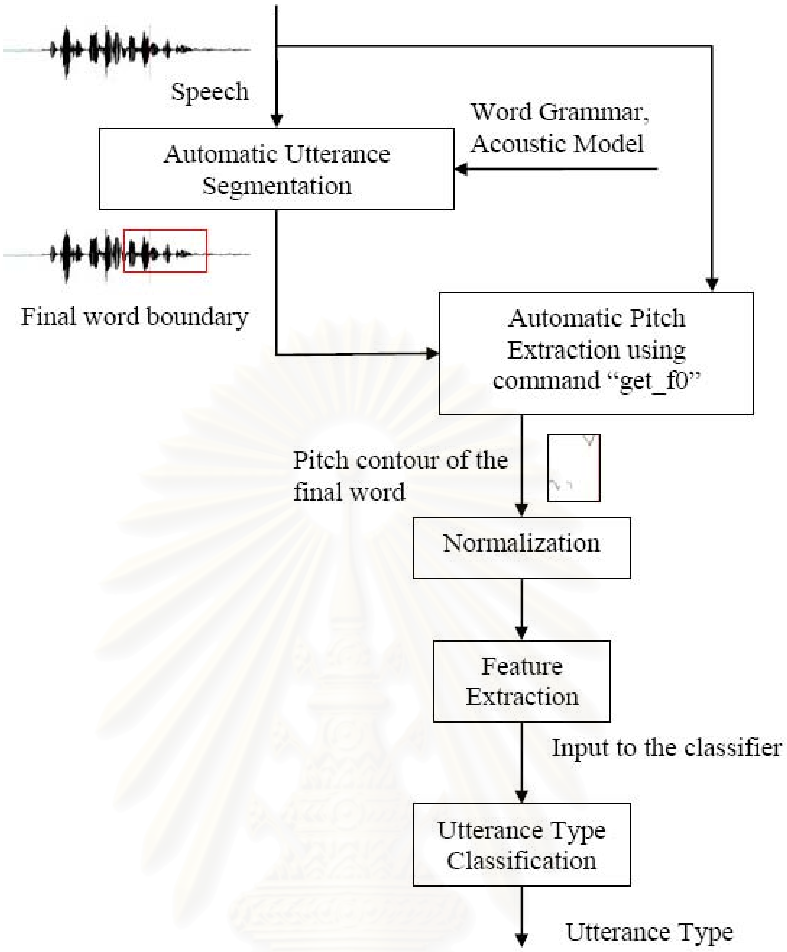


Figure 6.1 Block diagram of the automatic utterance-type recognizer

6.2.1 Automatic Utterance Segmentation

From several pairs of the utterances, it is found that the statements and declarative questions could be distinguished by listening only to their final words. On the basis of this finding, I decided to use the final word of each utterance as the data for utterance-type recognition.

To determine the final word boundary, an automatic utterance segmentation module is designed using the hidden Markov model toolkit (Young et al., 2002). I assumed that the transcription of each utterance is known, and performed alignment between each utterance and its transcription by using the Viterbi algorithm. For this procedure, I used a standard feature set in speech recognition and made an acoustic model using a relatively small amount of training data, which is different from the database described in the previous section. The detailed conditions will be explained in section 6.3.

6.2.2 F0 Extractor

The 'get_f0' program from the Entropic Waves software package (Entropic, 1993; Entropic, 1998) is used to extract *F0* data from the final word in each utterance. The ‘get_f0’ implements a fundamental frequency estimation algorithm using a normalized cross correlation function and a dynamic programming function. In the experiments described in the next section, the default values of the parameters of the 'get_f0', i.e., Gaussian window with the length of 0.04 sec, and the shift time of 0.01 sec are used. The *F0* data are converted into logarithmic scale and passed through a normalizer.

6.2.3 Normalizer

Even for the utterances of the same sentence with the same utterance type, the pitch contour of their final word may be different from speaker to speaker. This difference increases the variation in the pitch contour. To achieve robustness against this variation among speakers, the log*F0* values are normalized both in the frequency domain and in the time domain.

Let p_i (i = 1,2,...,L) and t_i (i = 1,2,...,L) be the sequences of the *logF0* values and the time of the final word with length *L*. The normalized vector of p_i, $\tilde{p}_i$ (i = 1,2,...,L) is calculated as

$$\tilde{p}_i = \frac{p_i - p_{\min}}{p_{\max} - p_{\min}}, \tag{6.1}$$

and the normalized vector of t_i, $\tilde{t}_i$ (i = 1,2,...,L) is calculated as

$$\tilde{t}_i = \frac{t_i - t_1}{t_L - t_1} \quad . \tag{6.2}$$

The normalized *logF0* values are input of the feature extractor.

6.2.4 Feature Extractor

Features of the *F0* contour are extracted using the polynomial expansion method (Levitt and Rabiner, 1971; Hwang and Chen, 1994), where the pitch contour is approximated as a polynomial line in two-dimensional plane of the normalized log*F0* and time. The coefficients c_i of the polynomial expansion are extracted using the Least Mean Square (LMS) algorithm.

Let N be the order of the polynomial expansion. Then the approximated contour for the normalized logF0, $\hat{\tilde{p}}$ can be expressed as

$$\hat{\tilde{p}} = \sum_{i=0}^{N} c_i \tilde{t}^{\,i} \quad . \tag{6.3}$$

The coefficient for i = 0 is removed to achieve robustness against the difference in the pitch level between male and female speakers.

Figure 6.2 shows the representation of typical F0 data using various order of polynomial expansion. The higher the order of the polynomial expansion was, the smaller the error between the estimated points with the original F0 data became. However, as will be explained in the next section, the best performance of the automatic utterance-type recognizer may not be achieved with the highest order of polynomial expansion because of the over-training problem.

The representation of the pitch contour of the final words in the proposed utterance-type recognizer will be compared with the representation of the pitch contour of the final words proposed by Ishi (2005) (Ishi, 2005).

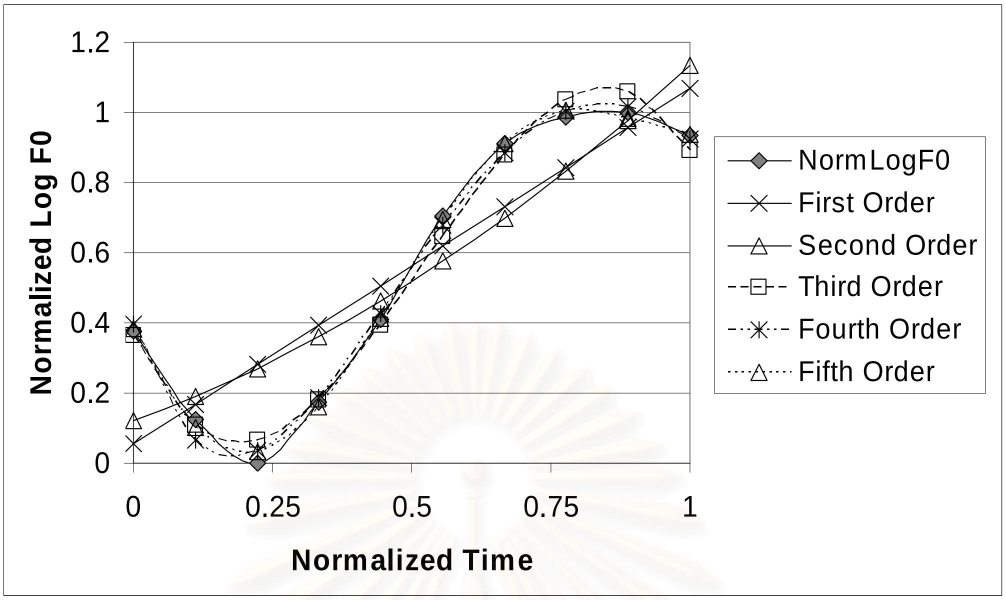


Figure 6.2 Representation of typical F0 data using various order of polynomial expansion

6.2.5 Classifier

A neural network (Katagiri, 2000) is used as a classifier in the automatic utterance-type recognizer. The neural network has three full-connected layers: the input, hidden, and output layers, as illustrated in Figure 6.3. The number of nodes in the input layer is equal to the number of features extracted from the *F0* contour. The number of nodes in the hidden layer is optimized through experiments. One node in the output layer, which is trained to output zero for the statement and one for the declarative question is used.

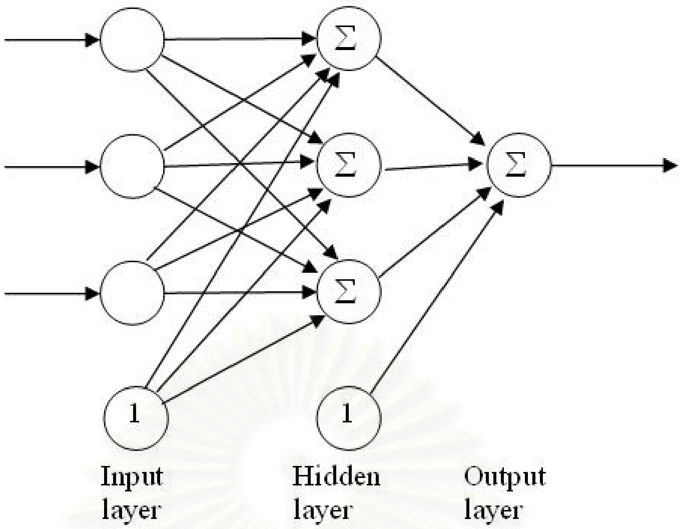


Figure 6.3 A three layer full-connected neural network

A statement-declarative question threshold is utilized at the end of the output node to classify the utterance type. The threshold is controlled in each experiment such that the error rates for both classes are equal.

Table 6.2 The four combinations of the training and the testing sets for the evaluation

Combination	Training Set	Testing Set
1	Set I	Set III
2	Set II	Set IV
3	Set III	Set I
4	Set IV	Set II

6.3 Experiments

6.3.1 Experimental conditions

For the automatic utterance segmentation described in subsection 6.2.1, an acoustic model is constructed in the following procedure. First, 3,840 utterances were recorded from 11 male and 9 female speakers, in which each speaker read texts from Indonesian newspapers and Indonesian linguistics books. The texts were chosen in order to contain all phonemes that appear in Indonesian speech. Then, for each 10 msec frame, features of the power and 12 mel-frequency cepstral coefficients (MFCC), and their first and second order derivatives were extracted. The total dimension of the feature vector was 39. Finally, using the training data, the monophone hidden Markov models (HMMs) with five states for each phone and 16 Gaussian mixtures for each state were trained.

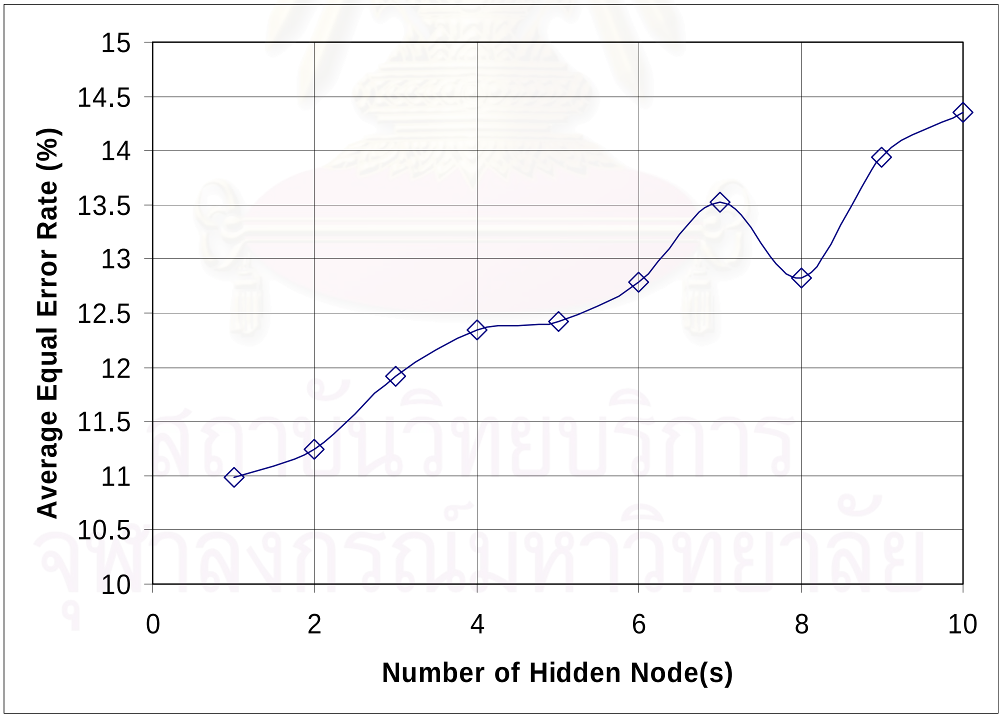


Figure 6.4 Equal Error Rate of the open test of the utterance-type recognizer using the third order of the polynomial expansion and various numbers of hidden nodes

Using the data sets of speech data as listed in Table 6.1, four combinations of training and testing sets were designed. The combinations are illustrated in Table 6.2. The classifier was trained for 100,000 epochs using a back-propagation algorithm.

Performance of the automatic utterance-type recognizer was evaluated by the averaged Equal Error Rate (EER), which is the average of the Equal Error Rates of the four combinations in Table 6.2.

6.3.2 Results

First, the number of hidden nodes in the neural network classifier was investigated. Figure 6.4 shows the EER of the utterance-type recognizer with various numbers of hidden nodes when the order of the polynomial expansion is three. The recognizer using one hidden node achieved the lowest EER. In this open test, the larger the number of hidden nodes, the larger the EER of the recognizer. The larger number of hidden nodes used in the neural networks means that the neural network will be more specific in learning the training set. Since the training set does not cover all variation of the pitch contour of the final word of speech data in the testing set, the more specific the neural network learn the training set, the larger error of the neural network in the recognition of the utterance type of the testing set. However, in the close test of the utterance-type recognizer, the higher the number of the hidden nodes, the smaller the EERs of the recognizer (see Figure 6.5).

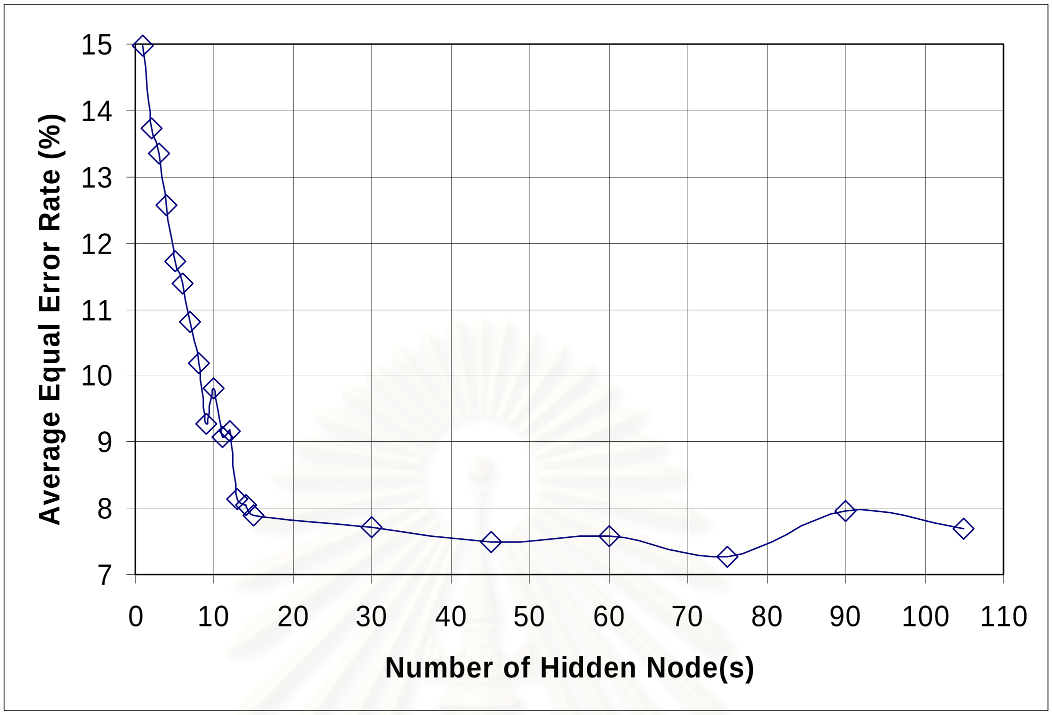


Figure 6.5 Equal Error Rate of the close test of the utterance-type recognizer using the third order polynomial expansion and various numbers of hidden nodes

Table 6.3 Equal Error Rate of the utterance-type recognizer using various order of the polynomial expansion and one hidden node

Order of the Polynomial Expansion	Equal Error Rate (%)
2	15.8
3	11.0
4	11.1
5	38.3

The utterance-type recognizer is assumed being used to recognize the utterance type of the testing set, which differs from the training set in both the sentence and the speaker. Therefore, based on the findings of the open test with the third order polynomial expansion, the performance of the utterance-type recognizers using the other orders of the polynomial expansion was investigated using one hidden node.

Table 6.3 shows the EERs of the utterance-type recognizer when the order of polynomial expansion is varied. The lowest EER was achieved when the third order polynomial expansion was used. The further increase of the order of the polynomial expansion increased the EER.

Next, the EERs of the proposed automatic utterance-type recognizer were compared with the EERs when the segmentation of the final words was carried out manually. The third order polynomial expansions were used and both recognizers used one hidden node. The recognizer with the manual segmentation achieved 88.1% accuracy, which is 0.9 point worse than the automatic recognizer did. When either the second or the fifth order polynomial expansion was used, the recognizer with the manual segmentation achieved accuracy of 86.3% and 67.1%, respectively, which are 2.1 and 5.4 point higher than the automatic recognizer did. When the fourth order polynomial expansion was used, the recognizer with manual segmentation achieved 86.8% accuracy, which is 2.1 point worse than the automatic utterance-type recognizer did. The average errors in the estimation of the beginning time and the duration of the final word were 43.5 msec and 75.4 msec, respectively. This small difference made the recognition rates of both the utterance-type recognizers comparable.

Next, the performance of the proposed automatic utterance-type recognizer was compared with the performance of the utterance-type recognizer using the features to represent the pitch contour of the final word proposed by Ishi (2005) (Ishi, 2005). EER of the utterance-type recognizer using F0Move1 = F0tgt2b - F0avg2a to represent the pitch contour of the final word is 13.3%. EER of the utterance-type recognizer using F0Move4 = F0tgt2b - F0tgt2a to represent the pitch contour of the final word is 14.5%. EER of the utterance-type recognizer using F0Move2 = F0avg2b - F0avg2a to represent the pitch contour of the final word is 40.5%. EER of the utterance-type recognizer using F0Move3 = F0avg2b - F0tgt2a to represent the pitch contour of the final word is 37.9%. The utterance-type recognizer using a combination of F0Move1 and F0Move4 with the neural networks consisting of two input nodes, six hidden nodes, and one output node achieved EER of 13.1%. Those experiments showed that the proposed utterance-type recognizer with the representation of pitch contour using the third order polynomial expansion is superior the utterance-type recognizer using the features proposed by Ishi (2005) to represent the pitch contour of the final word.

The average error rate of the proposed utterance-type recognizer is higher than that of the utterance-type recognizer based on Fujisaki model (Effendy et al., 2004). The recognizer in (Effendy et al., 2004) is speaker-dependent and evaluated using the testing set that covers the training set, and, therefore, difficult to be implemented as an automatic recognizer. On the other hand, the proposed recognizer is speaker-independent and evaluated using the testing set that differs from the training set in both the sentence and the speaker. Therefore, it is expected to be more robust in real application.

6.4 Summary

This chapter reports my study on an automatic utterance-type recognizer using the polynomial coefficients of the pitch contours of the sentence final words to identify declarative questions and statements in Indonesian speech. The automatic utterance-type recognizer consists of a pitch extractor, normalizer, feature extractor, classifier and an utterance segmentation module.

The findings of the study confirmed that the use of the final word of the utterance and the pitch contour information was effective in identifying the Indonesian declarative questions and statements. The highest recognition rate was achieved using the third order polynomial expansion as the feature extractor in the automatic utterance-type recognizer.

The proposed recognizer is speaker-independent and evaluated using the testing set that differs from the training set in both the sentence and the speaker. Consequently, it is expected to be more robust in real application.

CHAPTER VII
CONCLUSIONS

71 Conclusions of the Dissertation

This dissertation reports my study on the automatic utterance-type recognizer to identify the declarative questions and the statements in Indonesian speech.

At first, the utterance-type recognizer based on Fujisaki model has been designed to recognize the declarative questions and the statements in Indonesian speech. The utterance-type recognizer consists of a pitch extractor, a Fujisaki model, a Fujisaki-model parameter extractor, a Fujisaki-model parameter selector, and a classifier.

Three Fujisaki-model parameters are used in the dissertation: the amplitude of last accent command (A_{aJ}), (2) the magnitude of last phrase command ($A_{p\mathrm{I}}$), and (3) the baseline value of the fundamental frequency (F_b). Four different combinations of Fujisaki parameters were used as the input of the neural networks in the utterance-type recognizer: A_{aJ}; A_{aJ} and $A_{p\mathrm{I}}$; F_b, A_{aJ} and $A_{p\mathrm{I}}$; a fraction of F_b: F_b/100, A_{aJ} and $A_{p\mathrm{I}}$. The highest recognition rate of the utterance-type recognizer was achieved using a fraction value of F_b: F_b/100, the amplitude of last accent command, and the magnitude of last phrase command as the input of the neural networks in the utterance-type recognizer.

The study on the design of an Indonesian speech recognizer has been conducted to support the design of an automatic utterance-type recognizer. In the research, the higher the number of the Gaussian mixtures is used in the design of the Indonesian acoustic model, the higher both the percentage correct and the percentage accuracy of the speech recognizer are. The highest percentage correct and the highest percentage accuracy of the speech recognizer using one or two Gaussian mixtures were achieved using the combination of MFCCs with energy, and their first order derivatives. The highest percentage correct and the highest percentage accuracy of the speech recognizer using four, eight or 16 Gaussian mixtures were achieved using the combination of MFCCs with energy, their first and second order derivatives. In the research, the highest percentage correct and the highest percentage accuracy of the speech recognizer were achieved using the acoustic model created with 16 Gaussian mixtures and the combination of MFCCs with energy, their first and second order derivatives. This last combination was used to create the automatic utterance segmentation module of the automatic utterance-type recognizer.

The investigation of the characteristics of the final word of declarative questions and statements in Indonesian speech has been conducted. They include the pitch contour, pitch range, maximum pitch and the speech rate of the final words of the declarative questions and statements.

The average pitch range and the average maximum pitch of the final words of the declarative questions are larger than the average pitch range and the average maximum pitch of those of the statements. However, there is intersection of the distribution of the parameters. Consequently, an utterance-type recognizer may misrecognize the utterance type of some of the declarative questions and the statements when it uses only either the pitch range or the maximum pitch information.

Most of the final words of declarative questions were spoken at a slower rate than that of their corresponding statements. The duration per syllable of the final word of the statements and that of the declarative questions had a similar distribution. Therefore, the duration per syllable may be a low correlate to distinguish Indonesian statements and declarative questions.

The algorithm to extract the parameters of Fujisaki model is too complicated to be implemented in an automatic recognition system. Therefore, the utterance-type recognizer needs to be developed. an automatic utterance-type recognizer using the polynomial coefficients of the pitch contour of the sentence final words was proposed. The investigation in this dissertation confirmed that the use of the final word of the utterance and the pitch contour information were effective in the identification of the Indonesian declarative questions and statements. The highest recognition rate was achieved using the third order polynomial expansion as the feature extractor and a neural network with one hidden node as a classifier in the automatic utterance-type recognizer.

7.2 Contributions of the Dissertation

This section summarizes the contributions made during doing the research in this dissertation. The works begin with the design of the utterance-type recognizer for Indonesian speech based on Fujisaki model. I have been collected Indonesian speech data during the research. An Indonesian acoustic model has been design. Its performance as the part of an Indonesian speech recognizer and an Indonesian-utterance segmentation module of the automatic utterance-type recognizer has been investigated. The characteristics of the final words of Indonesian declarative questions and statements have been analyzed. They consist of the pitch contour, pitch range, maximum pitch and the speech rate of the sentence final words. The utterance-type recognizer has been developed to be an automatic system using the polynomial coefficients of the pitch contour of the sentence's final word. The details of the contributions are described as follows:

7.2.1 Indonesian speech data

Indonesian speech data are collected and recorded from 35 Indonesian native speakers. They consist of 11 female and 24 male speakers with age ranging from 23 to 50. The speech data are recorded in an office environment. The recorded speech data are digitized at 16 kHz sampling rate and 16-bit resolution.

The speech data are grouped into two domains: speech data A and speech data B. Speech data A are used to investigate the performance of the automatic utterance-type recognizer of Indonesian speech. Speech data B are used to create the Indonesian acoustic model used in the automatic utterance segmentation module of the automatic utterance-type recognizer.

To collect speech data A, the subjects are asked to utter 29 pairs of sentences as naturally as possible. The pair of the sentences consists of statements and their corresponding declarative questions. The sentences in the speech data A are with a different number of words. The sentences are selected from daily-life conversations among Indonesian speakers as listed in appendix A. From the recording, there are 2030 utterances in speech data A.

Speech data B are collected by recording the speakers reading out the Indonesian newspapers and Indonesian linguistics books. The texts are chosen in order to contain all phonemes that are possible in Indonesian speech.

Speech data A and speech data B are provided with their transcriptions. The transcriptions are used to label the speech data and to investigate the performance of the acoustic model used to create a speech recognizer or to create the utterance segmentation module of an automatic utterance-type recognizer.

7.2.2 Utterance-type recognizer for Indonesian speech based on Fujisaki model

This dissertation proposed an utterance-type recognizer for Indonesian speech based on Fujisaki model. The best recognition rate was achieved using a combination of a fraction of the baseline value of the fundamental frequency (F_b): F_b/100, the amplitude of last accent command, and the magnitude of last phrase command as the input of the neural networks in the utterance-type recognizer. To author’s knowledge, the utterance-type recognizer is the first one for Indonesian speech.

7.2.3 Characteristics of the final words of Indonesian declarative questions and statements

This dissertation revealed the characteristics of the final words of Indonesian declarative questions and statements. The characteristics comprise the pitch contour, pitch range, maximum pitch and the speech rate of the sentence’s final words. The characteristics may be utilized as the information to further develop the utterance-type recognizer.

7.2.4 Indonesian acoustic model

This dissertation provides an Indonesian acoustic model, which can be used to create an Indonesian speech recognizer or an Indonesian-utterance segmentation module of the automatic utterance-type recognizer. The highest percentage correct and the highest percentage accuracy of the speech recognizer were achieved using the acoustic model created with 16 Gaussian mixtures and the combination of MFCC, energy, their first and second order derivatives.

7.2.5 Automatic utterance-type recognizer using the polynomial coefficients of the pitch contour of the sentence's final word

This dissertation proposed an automatic utterance-type recognizer to distinguish the declarative questions and the statements of Indonesian speech using the polynomial coefficients of the pitch contour of the sentence final word. The automatic utterance-type recognizer may be combined with other spoken system to create a larger spoken system such as a spoken dialogue system or a spoken understanding system. The evaluation in this research confirmed that the use of the final word of the utterance and the pitch contour information was effective in the identification of the Indonesian declarative questions and statements.

7.2.6 Program Scripts Development

In this dissertation, some program scripts were created for the automatic pitch extractor, neural networks, Hidden Markov Model, and recognition rate calculation. The scripts can be used in the future researches of the study on the utterance-type recognition.

7.3 Future Research on the Utterance-type recognizer of Indonesian Speech

In the dissertation, the utterance-type recognizer was used to identify two utterance types: declarative questions and statements in Indonesian speech. In the evaluation of the automatic utterance-type recognizer, the training and the testing were set differently in terms of both the speakers and sentences. Patterns tested using the testing set may not be in the patterns of the training set. It made the recognizer misrecognized some patterns in the testing set. In the future, new larger speech data especially from female, which cover larger variation of the pitch contour of the utterances needs to be collected. Moreover, the automatic utterance-type recognizer will be developed to cover all types of Indonesian utterances including question, command, exclamation, and statement. To support the investigation, new speech data that cover all utterance types need to be collected. The automatic utterance-type recognizer may be able to be combined with other automatic system to create a larger spoken system such as a spoken dialogue system or a spoken understanding system to improve the performance of the spoken system.

When the Indonesian large vocabulary speech recognizer has been developed until it achieves its suitable performance, the automatic utterance segmentation module of the automatic utterance-type recognizer might be capable of being further developed. Therefore, the automatic segmentation module can segment more Indonesian sentences properly. Consequently, the automatic utterance-type recognizer will be capable of recognizing the utterance type of larger variation of the sentences. The automatic segmentation module may also be further developed by optimizing the parameters in HMM such as optimizing the number of states in HMM.

The methods used in this dissertation are expected to be also useful in the recognition of the utterance types of other similar languages such as Malay spoken in southern Thailand, Malaysia, and Brunei and Tagalog spoken in the Philippines.

REFERENCES

Adami, A. G., R. Mihaescu, D. A. Reynolds and J. J. Godfrey (2003). Modeling prosodic dynamics for speaker recognition. ICASSP 2003, Hongkong.

Aguero, P. D., K. Wimmer and A. Bonafante (2004). Automatic Analysis and Synthesis of Fujisaki's Intonation Model for TTS. Speech Prosody 2004, Nara, Japan.

Akagi, M. and T. Ienaga (1995). Speaker Individualities in Fundamental Frequency Contours and Its Control. EuroSpeech'95, Madrid, Spain.

Ali, A. M. A., J. V. de Spiegel and P. Mueller (2002). "Robust Auditory-Based Speech Processing Using the Average Localized Synchrony Detection." IEEE Transactions on Speech and Audio Processing **10**(5): 279-292.

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Oxford, England, Clarendon Press.

Boersma, P. and D. Weenink (2004). Praat: doing phonetics by computer, University of Amsterdam, http://www.praat.org.

Brickmann, C. and R. Benzmüller (1999). The Relationship between Utterance Type and F0 Contour in German. Eurospeech '99, Budapest, Hungary.

Brown-Schmidt, S., C. Gunlogson, D. Watson and M. T. Tanenhaus (2006). Perspective guides interpretation of questions, declarative questions and statements in unscripted conversation. The 10th Workshop on the Semantics and Pragmatics of Dialogue, University of Potsdam, Germany.

Bu, L. and T. D. Church (2000). "Perceptual Speech Processing and Phonetic Feature Mapping for Robust Vowel Recognition." IEEE Transactions on Speech and Audio Processing **8**(2): 105-114.

Carpenter, B. and J. Chu-Carrol (1999). Spoken Dialogue Systems, Lucent Technologies, Bell Labs Innovations.

Chen, K. and M. Hasegawa-Johnson (2005). "How Prosody Improves Word Recognition."

Dautrich, B., L. Rabiner and T. Martin (1983). "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition." IEEE Transactions on Acoustics, Speech, and Signal Processing **31**(4): 793-807.

Davis, S. B. and P. Mermelstein (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE Trans. Acoust. Speech Signal Processing **28**(4): 357-366.

Eady, S. J. and W. E. Cooper (1986). "Speech intonation and focus location in matched statements and questions." Journal of the Acoustical Society of America(80): 402-415.

Ebing, E. (1997). Form and Function of Pitch Movements in Indonesian. The Netherlands, Research School CNWS, Leiden University.

Effendy, N. (2002). Identification of the Frequency Spectrum of the Electrocardiograf Signals Using Full Competition Neural Networks. The Third National Seminar on Computational Intelligence, Jakarta, Indonesia.

Effendy, N., S. Hawibowo and B. Achmad (2001). "Optimization on Artificial Neural Networks to Control "kartini" Nuclear Research Reactor with Linear Reference Model." Media Teknik.

Effendy, N. and S. Jitapunkul (2006). Performance Evaluation of the Utterance Type Recognizer of Declarative Question and Statement in Indonesian Speech Based on Fujisaki Model. International Conference of the AUN/Seed-Net Field Wise Seminar on Multimedia Signal Processing and Communication System, Bangkok.

Effendy, N., S. Jitapunkul and S. Furui (2005). Combination of MFCC and Its Variances with Gaussian Mixture Incrementing in Speech Recognition. PACLING05, Tokyo.

Effendy, N., E. Maneenoi, P. Charnvivit, and S. Jitapunkul (2004). Intonation Recognition for Indonesian Speech Based on Fujisaki Model. Interspeech 2004, Korea.

Effendy, N., W. Setiawan and B. Achmad (1998). Implementation of Artificial Neural Networks on Frequency Spectrum Recognition. Seminar and Scientific Presentation of Basic Research in Nuclear Science and Technology, Yogyakarta, Indonesia, BATAN.

Entropic (1993). ESPS Version 5.0 Programs Manual. Washington, D.C., Entropic Research Laboratory.

Entropic (1998). ESPS Quick Reference. Washington, D.C., Entropic Research Laboratory, Inc.

Ezzaidi, H., and J. Rouat (2004). Pitch and MFCC dependent GMM models for speaker identification systems IEEE Canadian Conference on Electrical and Computer Engineering, Niagara Falls.

Fishel, M. (2006). Dialogue Act Recognition Techniques, in GSLT/NGSLT course on dialogue systems, Linkoping University, Sweden.

Fujisaki, H. and S. Ohno (1998). The Use of a Generative Model of F0 Contours for Multilingual Speech Synthesis. ICSP.

Fujisaki, H. and S. Ohno (1995). Analysis and Modelling of Fundamental Frequency Contours of English Utterances. Eurospeech'95, Madrid, Spain.

Fujisaki, H., S. Ohno and O. Tomita (1996). Automatic parameter extraction of fundamental frequency contours of speech based on a generative model ICSP.

Fujisaki, H. and H. Sudo (1971). "A model for the synthesis of prosodic pitch contours of connected japanese." Journal of the Acoustical Society of Japan **27**(8): 396-397.

Fujisaki, H., M. Tatsumi and N. Higuchi (1981). Analysis of pitch control in singing. Vocal Fold Physiology, University of Tokyo Press.

Furui, S. (2001). Digital Speech Processing, Synthesis, and Recognition, Mercel Dekker, Inc.

Grau, S., E. Sanchis, M. J. Castro and D. Vilar (2004). Dialogue Act Classification Using a Bayesian Approach. the 9th International Conference Speech and Computer (SPECOM’2004).

Gunlogson, C. (2001). True to Form: Rising and Falling Declaratives as Questions in English. Santa Cruz, CA, University of California. **Ph.D. Thesis**.

Haan, J., V. van Heuven, J. Pacilly and R. van Bezooijen (1997). Intonational Characteristics of Declarativity and Interrogativity in Dutch: A Comparison. ESCA workshop on intonation: Theory, Models, and Application, Athens, Greece.

Halim, A. (1981). Intonation in Relation to Syntax in Indonesian, Australian National University.

Higuchi, N., T. Hirai and Y. Sagisaka (1997). Effects of Speaking Style on Parameters of Fundamental Frequency Contour ‘Progress in Speech Synthesis. J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, Springer-Verlag**:** 417-428.

Hirst, D. J. and A. D. Cristo (1998). A survey of intonation systems. Intonation Systems : a Survey of Twenty Languages. D. J. Hirst and A. D. Cristo. Cambridge, Cambridge University Press**:** 1-44.

Hwang, S. H. and S. H. Chen (1994). Neural-network-based F0 text-to-speech synthesizer for Mandarin. Vis. Image Signal Process.

Ishi, C. T. (2005). "Perceptually-related F0 parameters for automatic classification of phrase final tones." IEICE Trans. Inf. & Syst. E88-D(3): 481-488.

Katagiri, S. (2000). Handbook of Neural Networks for Speech Processing. Boston, Artech House.

Keizer, S., R. op den Akker and A. Nijholt (2002). Dialogue Act Recognition with Bayesian Networks for Dutch Dialogues. 3rd SIGdial Workshop on Discourse and Dialogue, Philadelphia, PA, USA.

Kuwabara, H. and Y. Sagisaka (1995). "Acoustic Characteristics of Speaker Individuality : Control and Conversion." Speech Communication **16**: 165-173.

Ladd, D. R. (1996). Intonational Phonology. Cambridge, Cambridge University Press.

Lee, J., G. C. Kim and J. Seo (1997). A Dialogue Analysis Model with Statistical Speech Act Processing for Dialogue Machine Translation. the Spoken Language Translations EACL'97 Workshop, Budapest, Hungary.

Levin, L., C. Langley, A. Lavie, D. Gates, D. Wallace, and K. Peterson (2003). Domain Specific Speech Acts for Spoken Language Translation. the 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan.

Levitt, H. and L. R. Rabiner (1971). "Analysis of Fundamental Frequency Contour in Speech." The Journal of the Acoustical Society of America **49 (2)**: 569-582.

Maneenoi, E. (2004). An Acoustic Study Of Syllable Rhymes: A Basis For Thai Continuous Speech Recognition System. Bangkok, Thailand, Chulalongkorn University. **Ph.D. Thesis**.

Martin, T., T. Svendsen, and S. Sridharan (2003). Cross-Lingual Pronunciation Modelling for Indonesian Speech Recognition. Eurospeech, Geneva.

Matsumoto, H., Y. Nakatoh and Y. Furuhata (1998). An Efficient Mel-LPC Analysis Method for Speech Recognition. ICSLP.

Mixdorff, H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Istanbul, Turkey.

Mixdorff, H., H. Fujisaki, G. P. Chen and Y. Hu (2003). Towards the Automatic Extraction of Fujisaki model Parameters for Mandarin Eurospeech, Geneva.

Narusawa, S., N. Minematsu, K. Hirose and H. Fujisaki (2002). A Method for Automatic Extraction of Model Parameters from Fundamental Frequency Contours of Speech. ICASSP, Orlando, Florida.

Ngarmchatetanarom, N., E. Maneenoi, W. Asdonwised and S. Jitapunkul (2004). Tone Recognition of Thai Continuous Speech Using Fujisaki's Model. CCECE 2004 - CCGEI 2004, Niagara Falls, IEEE.

Odé, C. (1994). On the perception of prominence in Indonesian. Semaian 9: Experimental Studies of Indonesian Prosody. C. Odé and V. J. van-Heuven. Leiden, Department of Languages and Cultures of South-East Asia and Oceania, Leiden University**:** 27-107.

Odé, C. and V. J. van Heuven (1998). Word stress in Indonesian; its communicative relevance, Nijhoff KITLV Press.

Pierrehumbert, J. (1980). The Phonology and Phonetics of English Intonation Cambridge, MA, Ph.D. Dissertation, MIT.

Potisuk, S., M. P. Harper and J. Gandour (1999). "Classification of Thai Tone Sequences in Syllable-Segmented Speech Using the Analysis-by-Synthesis Method." IEEE Trans. Speech and Audio Processing **7**: 95-102.

Quinn, G. The Indonesian Language. http://www.hawaii.edu/sealit/Downloads/.

Rabiner, L. and B.-H. Juang (1993). Fundamentals of Speech Recognition. New Jersey, Prentice Hall PTR.

Reithinger, N. and E. Maier (1995). Utilizing statistical dialogue act processing in verbmobil. the 33rd Annual Meeting of the Association for Computational Linguistics.

Ries, K. (1999). HMM and Neural Network Based Speech Act Detection. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Phoenix, Arizona.

Rossi, P. S., F. Palmieri and F. Cutugno (2002). A Method for Automatic Extraction of Fujisaki-Model Parameters.

Sakti, S., A. A. Arman, S. Nakamura and P. Hutagaol (2004). Indonesian speech recognition for hearing and speaking impaired people. Interspeech, Jeju Island, Korea.

Samsuri (1978). Analisa Bahasa: Memahami bahasa secara ilmiah. Jakarta, Erlangga, Ltd.

Samuel, K., S. Carberry and K. Vijay-Shanker (1999). Automatically Selecting Useful Phrases for Dialogue Act Tagging. the 4th Conference of the Pacific Association for Computational Linguistics (PACLING'99), Waterloo, Ontario, Canada.

Silva, S. and S. Netto (2004). Closed-Form Estimation of the Amplitude Commands in the Automatic Extraction of the Fujisaki's Model. ICASSP 2004, Quebec, Canada.

Stack, M. (2005). Word Order and Intonation in Indonesian. WIGL.

Tan, J. Bahasa Indonesia: Between FAQs and Facts. http://www.indotransnet.com/article1.html.

van Heuven, V., J. Haan and R. S. Kirsner (1999). Phonetic correlates of sentence type in Dutch: Statement, question and command. ETRW on Dialogue and Prosody 1999, Veldhoven, The Netherland.

van Heuven, V. J. and E. van Zenten (2005). "Speech rate as a secondary prosodic characteristic of polarity questions in three languages." Speech Communication(47): 87-99.

Verhaar, J. W. M. (1992). Pengantar Linguistik (Introduction to Linguistics). Yogyakarta, Gadjah Mada University Press.

Wahlster, W., T. Bub and A. Waibel (1997). Verbmobil: The combination of deep and shallow processing for spontaneous speech translation. the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Munich, Germany.

Wang, C. (2001). Prosodic Modeling for Improved Speech Recognition and Understanding, MIT. **PhD Thesis**.

Watson, P. J. and D. Hughes (2006). "The Relationship of Vocal Loudness Manipulation to Prosodic F0 and Durational Variables in Healthy Adults." Journal of Speech, Language, and Hearing Research **49**: 636–644.

Wong, W., T. Martin, T. Svendsen and S. Sridharan (2003). Multilingual Phone Clustering for Recognition of Spontaneous Indonesian Speech Utilising Pronunciation Modeling Techniques. Eurospeech, Geneva.

Wright, H. (1998). Automatic Utterance Type Detection Using Suprasegmental Features. the 5th International Conference on Spoken Language Processing (ICSLP'98), Sydney, Australia.

Wright, H., M. Poesio and S. Isard (1999). Using High Level Dialogue Information for Dialogue Act Recognition Using Prosodic Features. an ESCA Tutorial and Research Workshop on Dialogue and Prosody, Eindhoven, The Netherland.

Xu, Y. (2005). "Speech melody as articulatorily implemented communicative functions." Speech Communication(46): 220-251.

Yan, Q., S. Vaseghi, D. Rentzos, C.-H. Ho and E. Turajlic (2003). Analysis of Acoustic Correlates of British, Australian and American Accents. ASRU 2003.

Young, S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland (2002). The Hidden Markov Model Toolkit [HTK] version 3.2.1 for Speech Processing, Cambridge University Engineering Department.

Yuan, J., C. Shih and G. P. Kochanski (2002). Comparison of Declarative and Interrogative Intonation in Chinese. Speech Prosody 2002, Aix-en-Provence, France.

APPENDICES

Appendix A

Indonesian Sentences Used to Create Pairs of Declarative Questions and Statements

Table A.1 Indonesian sentences used to create the pairs of declarative question and statement

(1) Tikar itu baru saja dicuci
[tikar itu baru saja dicuci]
plaited mat that just washed
That plaited mat was just washed

(2) Tembok itu dikotori oleh Iwan (person’s name)
[tembo’ itu dikotori oleh Iwan]
wall that durtied by Iwan
That wall is durtied by Iwan

(3) Jagoan itu menendang tiga orang penjahat
[Jagowan itu menendang tiga orang penjahat]
Hero that kick three (numeral classifier for humans) criminals
That hero kicks three criminals

(4) Polisi menangkap penjahat pagi tadi
[Polisi menangkap penjahat pagi tadi]
Policeman catch criminal morning this (just past)
Policeman catched a criminal this morning

(5) Sepeda iwan masih di bengkel
[Sepeda iwan masih di bengkel]
Bicycle Iwan still in garage
Iwan’s bicycle is still in a garage

(6) Dia sudah pergi tadi pagi
Dia sudah pergi tadi pagi
He already went this morning
He already went this morning

(7) Lisa (person’s name) sedang menyanyi dan menari
[Lisa sedang menyanyi dan menari]
Lisa is singing and dancing
Lisa is singing and dancing

(8) Dia sedang makan
[Dia sedang makan]
He is eating
He is eating

Table A.1 Indonesian sentences used to create the pairs of declarative question and statement (cont.)

(9) Albert (person’s name) lupa pada dirinya sendiri
[Albert lupa pada dirinya sendiri]
Albert forget about him self
Albert forget about himself

(10) kunci pintu itu dibobol maling
[kunci pintu itu dibobol maling]
key door that thief
That door is broken by a thief

(11) kucing telah menangkap seekor tikus
[kucing telah menangkap seekor tikus]
cat already catched a mouse
A Cat already catched a mouse

(12) Ibu sedang belanja ke pasar
[Ibu sedang belanja ke pasar]
mother is shopping in market
Mother is shopping in a market

(13) Jendela itu tidak bisa dibuka
[Jendela itu tidak bisa dibuka]
window that not can be opened
That window cannot be opened

(14) Headphone ini sangat bagus
[Headphone ini sangat bagus]
Headphone this very good
This headphone is very good

(15) Dokter sedang memeriksa pasien
[Dokter sedang memeriksa pasien]
Physician is checking patien
Physician is checking a patient

(16) Kursi ini baru dicat
[Kursi ini baru dicat]
chair this just painted
This chair is just painted

Table A.1 Indonesian sentences used to create the pairs of declarative question and statement (cont.)

(17) Pensil ini sudah tidak runcing
[pensil ini sudah tidak runcing]
pencil this already not sharp
This pencil is not sharp

(18) Gelas ini mudah pecah
[Gelas ini mudah pecah]
Glass this easily broken
This glass is easily broken

(19) Dia suka menolong orang lain
]Dia suka menolong orang lain]
He/she like help peope other
He/she likes to help other people

(20) Komputer itu terjangkit virus
[Komputer itu terjangkit virus]
Computer that infected virus
That computer is infected by virus

(21) Atap rumahnya sudah bocor
[Atap rumahnya sudah bocor]
Roof his house already
His house roof is already

(22) Lantai itu sudah kamu bersihkan
[Lantai itu sudah kamu bersihkan]
Floor that already you clean
The floor has been cleaned by you

(23) Adik sedang bermain
[Adik sedang bermain]
younger sister/brother is playing
The younger sister/brother is playing

(24) Andi berlari ke arah mobil
[Andi berlari ke arah mobil]
Andi run to car
Andi run to a car

Table A.1 Indonesian sentences used to create the pairs of declarative question and statement (cont.)

(25) Kucing dan anjing sedang berkelahi
[Kucing dan anjing sedang berkelahi]
cat and dog are fighting
A cat and a dog are fighting

(26) Televisimu telah dibeli oleh Umar
[Televisi mu telah dibeli oleh Umar]
television your has been bought by Umar (people's name)
Your television has been bought by Umar

(27) Samsul belum selesai membaca buku itu
[Samsul belum selesai membaca buku itu]
Samsul (person's name) not yet finish read book that
Samsul has not read the book yet

(28) Bapak sedang mengajar matematika
[Bapak sedang mengajar matematika]
Father is teaching mathematics
Father is teaching mathematics

(29) Air di tangki sudah penuh
[Air di tangki sudah penuh]
water in tank already full
The water in the tank is already full

Appendix B

Group of Speech Data for the Testing and the Training of Utterance-type recognizer using the Polynomial Coefficients of the Pitch Contours of Sentence Final Words

Table A.2 The set of the speech data for the investigation of the performance of the automatic utterance-type recognizer

		Speakers											
		F1, **M1**,**M7**,**M15** F4,F6,F8,F10,M4,M6, M10,M12,M14,M18,M20,M22,M24						**F2,M2, M8,M16** F3,F5,F7,F9,F11,M3,M5, M9,M11,M13,M17,M19,M21, M23					
		statement		fall declarative question		Rise declarative question		statement		fall declarative question		Rise declarative question	
		M	F	M	F	M	F	M	F	M	F	M	F
Sentences		**Set 1**						**Set 2**					
	1. air	12	5		1	12	4	12	6	1		11	6
	2. baca	12	5		1	12	4	12	6	1		11	6
	3. beli	10	5	2	1	10	4	12	6		1	12	5
	4. berkelahi	11	5		1	12	4	12	6	1		11	6
	5. berlari	12	5	1		11	5	12	6	2		10	6
	6. bermain	12	5	1		11	5	12	6	1		11	6
	7. bersih	11	5		1	11	4	12	6	1		11	6
	8. bocor	12	5	2	1	10	4	12	6	1		11	6
	9. cat	12	5	2	1	10	4	12	6	1	1	11	5
	10. dokter	12	5	1	1	11	4	12	6	1	2	11	4
	11. headphone	12	5		1	12	4	11	6	1		10	5
	12. jendela	12	5		1	12	4	12	6	1		10	6
	13. kepasar	12	5	1		11	5	12	6	1		11	6
	14. komputer	11	5		1	12	4	12	6	1		10	6
	subtotal	**163**	**70**	**10**	**11**	**157**	**59**	**167**	**84**	**14**	**4**	**151**	**79**
		Set 4						**Set 3**					
	1. kucing	12	5	3	1	9	4	11	6	2		9	5
	2. kunci	10	5		1	12	4	12	6	1		11	6
	3. lupa	12	5			12	4	12	6	1		11	6
	4. makan	12	5	2	1	10	4	12	6	1	1	10	5
	5. matematika	12	5		1	12	4	12	6	1		11	6
	6. menari	12	5	1	1	11	4	12	6	1		11	6
	7. menolong	11	5		1	12	3	12	6	1		10	6
	8. pecah	11	5		1	12	4	12	6	1		11	6
	9. pensil	12	5	2	1	10	4	12	6	1	1	11	5
	10. pergi	12	5		1	12	4	12	6	1	2	11	4
	11. sepeda	12	5	3	1	8	4	12	6	2	2	11	4
	12. tangkap	11	5	1	1	10	4	12	6	1		11	5
	13. tembok	12	5			12	5	12	6	1		11	6
	14. tendang	11	5			11	4	12	6	1		11	6
	15. tikar	11	5			12	5	12	5	1	1	11	5
	Subtotal	**173**	**75**	**12**	**11**	**165**	**61**	**179**	**89**	**17**	**7**	**161**	**81**

Table A.3 Detail of the fall declarative questions in the set of the speech data for the investigation of the performance of the automatic utterance-type recognizer

		Speakers										
		F1, **M1**, **M7**,**M15** F4,F6,F8,F10,M4,M6,M10,**M12**,M14,M18,M20,M22,M24					**F2**,**M2**, **M8**,**M16** F3,F5,F7,F9,F11,M3,M5,M9,M11,M13,M17,M19, M21, M23					
		fall declarative question					fall declarative question					
		M1	M6	M10	M15	F10	M3	M8	M9	M16	F2	F9
Sentences		**Set 1**					**Set 2**					
	1. air					1	1					
	2. baca					1	1					
	3. beli	1		1		1						1
	4. berkelahi					1	1					
	5. berlari		1				1		1			
	6. bermain			1			1					
	7. bersih					1	1					
	8. bocor	1		1		1	1					
	9. cat	1	1			1	1				1	
	10. dokter			1		1	1				1	1
	11. headphone					1	1					
	12. jendela					1	1					
	13. kepasar				1		1					
	14. komputer					1	1					
	subtotal	**3**	**2**	**4**	**1**	**11**	**13**		**1**		**2**	**2**
		10				**11**	**14**				**4**	
		Set 4					**Set 3**					
	1. kucing	1		1	1	1	1	1				
	2. kunci					1	1					
	3. lupa						1					
	4. makan	1		1		1	1				1	
	5. matematika					1	1					
	6. menari				1	1	1					
	7. menolong					1	1					
	8. pecah					1	1					
	9. pensil	1			1	1	1					1
	10. pergi					1	1					2
	11. sepeda	1		1	1	1	1			1	1	1
	12. tangkap	1				1	1					
	13. tembok						1					
	14. tendang						1					
	15. tikar						1				1	
	Subtotal	**5**		**3**	**4**	**1** **1**	**15**	**1**		**1**	**3**	**4**
		12				**11**	**17**				**7**	

Appendix C

Transcription of Labeled Indonesian Speech Data

1. Aesop4_02_eff.wav

konsonan ada yang bersuara, yang terjadi bila ada alur sempit di antara pita suara dan ada yang tidak bersuara, yang terjadi bila tempat artikulasi yang bersangkutan sajalah yang merupakan alur sempit, sedang pita suara itu terbuka agak lebar.

2. Aesop4_03_eff.wav

Untuk menjelaskan fungsi pita pita suwara, di bawah ini dimuat gambar ke empat posisi pita pita tersebut

3. Aesop4_04_eff.wav

udara dipompakan dari paru paru, melalui batang tenggorokan ke pangkal tenggorok, yang di dalamnya terdapat pita pita suara

4. Aesop4_05_eff.wav1

Pita suara itu harus terbuka, untuk memungkinkan arus udara keluar melalui rongga mulut, melalui rongga hidung atau melalui kedua duanya

5. Aesop4_06_eff.wav

fonitik auditoris tidak banyak dikerjakan dalam hubungan dengan linguistik. Buku-buku standar mengenai linguistik juga sedikit sekali menguraikan mengenai fonitik auditoris itu. Dan keahlian yang dituntut sebenarnya, adalah keahlian dalam ilmu kedokteran

6. Aesop4_07_eff.wav

yang ketiga adalah fonetik organis. Fonetik organis menyelidiki bagaimana bunyi-bunyi bahasa dihasilkan dengan alat-alat bicara. Bidang itu penting sekali untuk linguistik dan akan kita bicarakan secara terperinci dalam bab ini

7. Aesop4_08_eff.wav

Dalam bagian-bagian yang berikut, fonetik auditoris tidak diuraikan lagi. Hanya beberapa fakta dari fonetik akustis akan diuraikan, sehubungan dengan perlunya keterangan minimal demi uraian fonetik organis

8. Aesop4_09_eff.wav

hal pertama yang perlu diuraikan dalam fonetik organis ialah alat-alat bicara. Gambar yang berikut dengan daftar nama alat-alat tersebut kiranya cukup memadai

9. Aesop4_10_eff.wav

selaras dengan uraian tentang sistematik bahasa itu tadi, bab tiga membicarakan fonetik, bab empat membicarakan kronologi, bab lima membicarakan morfologi dan bab enam membicarakan sintaksis

10. Aesop4_11_eff.wav
linguistik sebagai ilmu pengetahuan, membutuhkan suatu teori yang konsekwen, sesuatu teori linguistis. Bila seorang ahli linguistik memusatkan perhatiannya khusus pada pendirian sesuatu teori, maka apa yang dikerjakannya boleh disebut linguistik teoritis.

11. Aesop4_12_eff.wav
semantik adalah cabang sistematik bahasa, yang menyelidiki makna atau arti. Seperti sudah dicatat di atas, perbedaan di antara leksikon dan gramatika, menyebabkan bahwa dalam semantik itu kita bedakan pula, antara semantik leksikal dan semantik gramatikal

12. Aesop4_13_eff.wav
Menurut sistematiknya, dalam setiap bahasa dapat dibedakan antara tata bahasa atau gramatika bahasa itu, dan perbendaharaan kata atau leksikon dalam bahasa yang sama. Oleh sebab itu, analisa tata bahasa atau analisa gramatikal dibedakan dari analisa leksikon, atau leksikologi atau analisa leksikal

13. Aesop4_14_eff.wav
dalam pengantar satu ini, yang sekiranya akan menarik perhatian ialah, bahwa banyak hal diandaikan kebenarannya tanpa bukti, atau tanpa bukti lengkap. Memang hal itu biasa terjadi dalam sebuah buku dikdaktis dan tidak jarang terjadi bahwa, bukti tentang hal-hal yang elementer menuntut keahlian tinggi sehingga jelas tidak dapat dimasukkan dalam sebuah buku pegangan

14. Aesop4_15_eff.wav
akhirnya saya dengan senang hati ingin memenuhi kewajiban mengucapkan rasa terima kasih dan penghargaan kepada mereka yang telah membantu dalam penyusunan buku ini. beberapa draft pertama untuk bab tiga sampai dengan lima pernah disusun oleh asisten saya di ui, dokterhandes el sihombing

15. Aesop5_01_eff.wav
di pihak lain, rupa-rupanya ada masalah yang lain, yang akan memunculkan pertanyaan atau kritik. Pendirian yang terdapat dalam buku ini, rupa-rupanya, dapat memberi kesan agak artikuler penentuannya

16. Aesop5_02_eff.wav
yang hendak saya jelaskan sebagai masalah yang pertama, ialah seleksi pendekatan pada umumnya, karena harus ada prinsip priyoritas. Prinsip semacam itu, dapat dibagi atas yang negatif dan yang positif

17. Aesop5_03_eff.wav
seperti diketahui oleh semua dosen yang pernah mencoba menyusun buku pegangan, tugas itu boleh dikatakan menuntut banyak, antara lain justru karena bahannya sederhana. Bila dipandang dari sudut keahlian profesional // tugas ini semakin berat. Saya pun dalam penyusunan buku pengantar linguistik ini telah mengalami kesulitan-kesulitan semacam itu

18. Aesop5_04_eff.wav
kekeliruan-kekeliruan yang masih ada menjadi tanggung jawab saya seluruhnya. Saya akan menyambut segala kritik dengan senang hati, agar supaya buku ini di kemudian hari dapat disempurnakan.

19. Aesop5_05_eff.wav
akhirnya, cukup banyak orang lain yang juga membantu saya. Beberapa rekan dosen memberi catatan dan kritik. Mahasiswa-mahasiswa yang sering tanpa menyadarinya // menyebabkan saya memikirkan kembali beberapa bahan.

20. Aesop5_06_eff.wav
dewasa ini, fidologi diartikan sebagai ilmu yang menyelidiki masa kuno, dari sesuatu bahasa berdasarkan dokumen-dokumen tertulis. Walaupun para ahli fidologi sekarang, menyadari bahwa pengetahuwan sedikit tentang linguistik dapat menjadi panduan penting dalam bidang mereka, namun sudahlah menjadi pengertian bersama // bahwa fidologi tidak sama dengan linguistik

21. Aesop5_07_eff.wav
pada umumnya // yang dimaksudkan dengan distribusi ialah kemungkinan penggantian konstituwen tertentu dalam kalimat tertentu dengan konstituwen yang lain. misalnya dalam kalimat tadi, konstituwen putri bisa diganti dengan konstituwen putra atau anak atau lurah, tapi bukannya dengan berjalan, atau sering atau aduh.

22. Aesop5_08_eff.wav
dalam bab enam sudah diuraikan gejala-gejala yang terpenting, yang berhubungan dengan pembedaan antara fungsi, kategori dan peran sintaksis. Bidang sintaksis begitu luas sehingga soal-soalnya tak kunjung habis dapat ditambahkan

23. Aesop5_09_eff.wav
memang seringkali pemakaian kata “adalah” itu tidak perlu, misalkan pada kalimat tadi, dapat berbunyi juga itu tidak benar. namun, seringkali adapula pemakaian kata adalah yang tidak mencerminkan pengaruh barat atau mencerminkan pemakaian motif kata kepula dalam bahasa-bahasa indo eropa

24. Aesop5_10_eff.wav
kata kerja transitif dalam arti tradisional istilah tersebut, sering dipakai tanpa objek, misalkan pada kata kerja ‘makan’ dalam kalimat ‘saya sudah makan’. Pemakaian kata kerja transitif demikian, disebut pemakaian yang absolut. Istilah tersebut berasal dari kata latin absolutus, yang berarti dilepaskan atau terlepas

25. Aesop5_11_eff.wav
sebagai masalah terakhir harus kita uraikan sedikit tentang hubungan antara fungsi dan peran.

26. Aesop5_12_eff.wav
di pihak lain, pasti ada peran bawahan dalam arti, terdapat dalam frase. Dalam frase ‘rumah tetangga saya’, konstituwen ‘tetangga saya’ berperan posisi, sedangkan dalam frase ‘pengeluaran uang’, konstituwen ‘uang’ berperan objektif karena uangnya dikeluarkan tidak mengeluarkan sesuatu

27. Aesop5_13_eff.wav
teori peran masih berada dalam tahap primitif. Istilah-istilah yang disebutkan di atas tidak semua dicontohkan di sini karena banyak dapat dipersoalkan. Kita dapat mencari contoh-contoh sendiri, tidak usah kita garap semuanya, sudah cukup bila beberapa istilah sudah pernah kita temui

28. Aesop5_14_eff.wav
kesimpulannya jelas, dalam banyak bahasa astronesia seperti bahasa tagalog, indonesia dan jawa, struktur peran menyebabkan bentuk kategorial di tempat predikat menyesuaikan diri dengan peran yang terdapat di tempat subjek.

29. Aesop5_15_eff.wav
soalnya tidak menyangkut hanya bentuk pasif tadi saja, karena baik kata benda maupun kata kerja, meskipun kedua kategori terdapat baik dalam bahasa-bahasa astronesiya maupun dalam bahasa-bahasa indo eropa, tidak mutlak perlu persis sama dengan kedua tipe bahasa tersebut

30. Aesop5_16_eff.wav
fungsi-fungsi itu sendiri tidak memiliki bentuk tertentu, tetapi harus diisi oleh bentuk tertentu, yaitu suatu kategori

5_16b
fungsi-fungsi itu juga tidak memiliki makna tertentu, tetapi harus diisi oleh makna tertentu yaitu peran

31. Aesop7_01_eff.wav
yang menarik perhatian bila kita bandingkan hasil pertama dengan lajur ke empat di atas ialah bahwa dalam lajur pertama kita pakai istilah asli indonesia sedangkan bentuk aktif tifalnya dalam lajur keempat merupakan kata pinjaman asing. Fenomena itu dalam bahasa indonesia memang tidak terbatas pada peristilahan ilmu linguistik saja.

32. Aesop7_02_eff.wav
maka tidaklah mengherankan apabila centang perentang politik di indonesia saat ini disebabkan oleh tiadanya hukum yang andal dan tangguh untuk membatasi, mengerem dan memaksa politik untuk mengikuti trek hukum

33. Aesop7_03_eff.wav
Padahal hukum bukanlah atau seharusnya bukanlah seperti yang diungkapkan di muka, tetapi hukum adalah sebagai alat mencapai keadilan. Bila hukum dengan ayat, pasal dan dalil-dalilnya penuh dengan multitafsir, maka jebol dan tumbanglah makna dan benteng keadilan

34. Aesop7_04_eff.wav
Bila keberadaan hukum di Indonesia berlanjut seperti ini maka yang diharapkan oleh rakyat bukanlah sekedar low imforsment yang pada akhirnya juga tak akan ada faedahnya, tetapi yang lebih menguatirkan terjadinya lowfus stet

35. Aesop7_05_eff.wav
Maka janganlah disalahkan apabila masyarakat yang sudah buta hukum tetapi sedang mendambakan keadilan berperasangka bahwa dalam peradilan tahu-tahu tujuannya atau keputusannya telah ditentukan lebih dahulu, barulah kemudian dicarikan dalil-dalil pembenaran hukumnya

36. Aesop7_06_eff.wav
pemerintah mengakui jumlah penderita demam berdarah mengalami peningkatan dua kali lipat perbulan dibandingkan tahun lalu

37. Aesop7_07_eff.wav
Jika tahun lalu dalam setahun tercatat lima puluh ribu dua puluh lima penderita, ini berarti tiap bulan ada empat puluh ribu orang penderita

38. Aesop7_08_eff.wav
menko kesra Yusuf kalla mengatakan masalah demam berdarah lebih berbahaya ketimbang sars dan flu burung karena demam berdarah sudah menimbulkan korban jiwa begitu banyak

39. Aesop7_09_eff.wav
oleh karena itu baik pemerintah dan masyarakat, menurut kalla // harus menyelesaikan masalah ini secepatnya. pasalnya, akibat demam berdarah akan jauh lebih besar.

40. Aesop7_10_eff.wav
kalla menambahkan penanganan demam berdarah harus melibatkan seluruh masyarakat, tidak cukup kalau ditangani pemerintah saja

41. Aesop7_11_eff.wav
selain itu, dalam waktu dekat pihak kesra akan mengumpulkan para gubernur, bupati dan walikota guna membahas keterlibatan masyarakat menanggulangi demam berdarah bersama. Kita akan menggerakkan masyarakat agar terlibat dalam perang terhadap nyamuk // untuk mencegah demam berdarah

42. Aesop7_12_eff.wav
ketua panitia pelaksana // haji abdul latif usman ketika dihubungi di balik papan jumat mengatakan pihaknya sudah menghubungi keduanya di Jakarta dan apabila sesuai rencana // kedua tokoh nasional itu dipastikan hadir pada acara tersebut

43. Aesop7_13_eff.wav
dijelaskan bahwa peserta es en je dua ribu empat // diperkirakan akan dihadiri oleh dua ratus da 'i // yang merupakan utusan dewan suroh lain daerah hidayatullah seIndonesia, utusan ormas, bakor islam dan lembaga dakwah islam

44. Aesop7_14_eff.wav
saat ini sudah hadir lebih dari enam puluh orang utusan Depede hidayattullah. Diharapkan pada jum'at dan sabtu, seluruh peserta bisa hadir seluruhnya, katanya

45. Aesop7_15_eff.wav
acara es 'ende dua ribu empat bertujuan untuk menata kembali langkah dan program dakwah

Appendix D

Error Rate of the Automatic Utterance-type recognizer

Table A.4 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-1-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	10.41	10.41	10.41
IV	II	10.68	10.26	10.47
I	III	10.37	10.79	10.58
II	IV	12.66	12.24	12.45
Average		11.03	10.92	10.98

Table A.5 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-2-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	10.41	9.95	10.18
IV	II	11.54	11.54	11.54
I	III	10.79	10.79	10.79
II	IV	12.24	12.66	12.45
Average		11.24	11.23	11.24

Table A.6 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture of the neural networks 3-3-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	10.86	10.41	10.63
IV	II	11.97	12.39	12.18
I	III	12.86	12.86	12.86
II	IV	11.81	12.24	12.03
Average		11.87	11.97	11.92

Table A.7 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-4-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	14.48	11.76	13.12
IV	II	13.68	13.68	13.68
I	III	10.37	10.37	10.37
II	IV	11.81	12.66	12.24
Average		12.58	12.12	12.35

Table A.8 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-5-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	11.31	10.41	10.86
IV	II	14.11	14.53	14.32
I	III	11.62	12.03	11.83
II	IV	12.66	12.66	12.66
Average		12.42	12.41	12.42

Table A.9 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-6-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	10.41	10.86	10.63
IV	II	15.38	15.81	15.6
I	III	11.62	11.62	11.62
II	IV	13.08	13.51	13.29
Average		12.62	12.95	12.78

Table A.10 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-7-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	11.31	10.86	11.09
IV	II	15.81	16.67	16.24
I	III	13.69	13.69	13.69
II	IV	13.08	13.08	13.08
Average		13.47	13.57	13.52

Table A.11 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-8-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	10.41	10.41	10.41
IV	II	14.11	14.53	14.32
I	III	12.45	12.03	12.24
II	IV	14.35	14.35	14.35
Average		12.83	12.83	12.83

Table A.12 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-9-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	14.03	14.03	14.03
IV	II	15.81	15.81	15.81
I	III	12.86	12.86	12.86
II	IV	13.08	13.08	13.08
Average		13.94	13.94	13.94

Table A.13 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-10-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	13.12	13.57	13.35
IV	II	16.24	16.67	16.45
I	III	14.11	14.11	14.11
II	IV	13.51	13.51	13.51
Average		14.24	14.46	14.35

Table A.14 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-11-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	13.57	12.22	12.9
IV	II	17.09	17.09	17.09
I	III	14.94	14.94	14.94
II	IV	14.35	15.19	14.77
Average		14.99	14.86	14.92

Table A.15 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-12-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	13.57	13.57	13.57
IV	II	16.24	16.24	16.24
I	III	12.45	12.45	12.45
II	IV	15.19	15.19	15.19
Average		14.36	14.36	14.36

Table A.16 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-13-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	15.84	15.38	15.61
IV	II	17.52	18.38	17.95
I	III	12.86	12.86	12.86
II	IV	13.51	13.92	13.71
Average		14.93	15.13	15.03

Table A.17 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-14-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	14.03	14.03	14.03
IV	II	17.95	14.96	16.45
I	III	12.86	12.86	12.86
II	IV	15.19	14.77	14.98
Average		15.01	14.15	14.58

Table A.18 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-15-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	14.03	14.03	14.03
IV	II	17.95	18.38	18.16
I	III	11.62	15.35	13.49
II	IV	10.97	11.81	11.39
Average		13.64	14.89	14.27

Table A.19 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-30-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	11.76	12.22	11.99
IV	II	16.67	17.09	16.88
I	III	13.28	13.69	13.49
II	IV	14.77	15.19	14.98
Average		14.12	14.55	14.33

Table A.20 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-45-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	14.03	14.93	14.48
IV	II	17.52	17.52	17.52
I	III	13.69	14.11	13.91
II	IV	13.51	13.92	13.71
Average		14.69	15.12	14.9

Table A.21 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-60-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	14.48	14.03	14.25
IV	II	14.53	14.96	14.74
I	III	13.69	13.69	13.69
II	IV	15.19	15.61	15.41
Average		14.47	14.57	14.52

Table A.22 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-75-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	14.93	15.38	15.16
IV	II	15.38	15.81	15.6
I	III	12.86	14.52	13.69
II	IV	15.61	16.03	15.82
Average		14.69	15.43	15.07

Table A.23 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-90-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	13.57	13.57	13.57
IV	II	16.24	16.67	16.45
I	III	9.13	14.94	12.03
II	IV	14.35	14.77	14.56
Average		13.32	14.99	14.15

Table A.24 Error rate of the automatic utterance-type recognizer using the third order polynomial expansion, architecture 3-105-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	14.48	14.93	14.71
IV	II	17.09	17.95	17.52
I	III	9.54	14.52	12.03
II	IV	15.19	16.03	15.61
Average		14.07	15.86	14.97

Table A.25 Error rate of the automatic utterance-type recognizer using the second order polynomial expansion, architecture 2-1-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	14.9	14.9	14.9
IV	II	15.4	14.5	14.9
I	III	18.3	17.4	17.8
II	IV	15.2	16.0	15.6
Average		15.9	15.7	15.8

Table A.26 Error rate of the automatic utterance-type recognizer using the fourth order polynomial expansion, architecture 4-1-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	10.9	10.9	10.9
IV	II	10.7	10.3	10.5
I	III	10.8	11.2	11.0
II	IV	11.8	12.2	12.0
Average		11.0	11.1	11.1

Table A.27 Error rate of the automatic utterance-type recognizer using the fifth order polynomial expansion, architecture 5-1-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	41.2	35.3	38.2
IV	II	39.3	38.0	38.7
I	III	43.9	31.5	37.8
II	IV	46.8	29.9	38.4
Average		42.8	33.7	38.3

Appendix E

Error Rate of the Automatic Utterance-type recognizer with Manual Utterance Segmentation

Table A.28 Error rate of the semi automatic utterance-type recognizer using the second order polynomial expansion, architecture 2-1-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	9.51	11.31	10.41
IV	II	14.53	15.38	14.96
I	III	14.52	14.52	14.52
II	IV	14.77	14.77	14.77
Average		13.33	13.99	13.66

Table A.29 Error rate of the semi automatic utterance-type recognizer using the third order polynomial expansion and architecture 3-1-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	9.5	9.9	9.7
IV	II	11.5	12.4	11.7
I	III	12.0	12.0	12.0
II	IV	14.4	14.4	14.4
Average		11.9	12.2	11.9

Table A.30 Error rate of the semi automatic utterance-type recognizer using the fourth order polynomial expansion and architecture 4-1-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	15.4	7.2	11.3
IV	II	14.5	14.1	14.3
I	III	10.4	16.2	13.3
II	IV	10.6	16.5	13.5
Average		12.7	13.5	13.1

Table A.31 Error rate of the semi automatic utterance-type recognizer using the fifth order polynomial expansion and architecture 5-1-1

Training Set	Testing Set	Error Rate (%)		Equal Error Rate (%)
		Statement	Declarative Question	
III	I	48.4	28.9	38.7
IV	II	14.5	42.7	28.6
I	III	44.0	22.8	33.4
II	IV	43.9	18.1	31.0
Average		37.7	28.2	32.9

Vitae

Nazrul Effendy received the ir. (Sarjana Teknik) degree in instrumentation engineering and M.Eng. degree in Electrical Engineering from Gadjah Mada University, Indonesia in 1998 and 2001, respectively. He has been a faculty staff at the Department of Engineering Physics, Gadjah Mada University since 1998. Since November 2003, he has been doing a Ph.D. program at Chulalongkorn University, Thailand. From September 13, 2005 to August 2, 2006, he conducted research about utterance-type recognizer at Furui Laboratory, Department of Computer Science, Tokyo Institute of Technology, Japan. His research interests include speech signal processing, prosody, and pattern recognition. He is a member of ISCA.