

UNIVERSITI TEKNOLOGI MARA

**DISCRIMINATIVE CLASSIFICATION MODEL
OF FILLED PAUSE AND ELONGATION FOR
MALAY LANGUAGE SPONTANEOUS SPEECH**

RASEEDA BINTI HAMZAH

PhD

April 2016

UNIVERSITI TEKNOLOGI MARA

**DISCRIMINATIVE CLASSIFICATION MODEL
OF FILLED PAUSE AND ELONGATION FOR
MALAY LANGUAGE SPONTANEOUS SPEECH**

RASEEDA HAMZAH

Thesis submitted in fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Computer and Mathematical Science

April 2016

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a panel of examiners has met on 20th January 2016 to conduct the final examination of Raseeda Binti Hamzah on his Doctor of Philosophy thesis entitled “Discriminative Classification Model of Filled Pause and Elongation for Malay Language Spontaneous Speech” in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiners recommends that the student be awarded the relevant degree. The panel of Examiners was as follows:

Puzziawati Ab.Ghani, PhD
Associate Profesor
Faculty of Computer and Mathematical Science
Universiti Teknologi MARA
(Chairman)

Ismail Musirin, PhD
Professor
Faculty of Electrical Engineering
Universiti Teknologi MARA
(Internal Examiner)

Syed Abdul Rahman Al-Haddad Syed Mohamed, PhD
Associate Professor
Faculty of Engineering
Universiti Putra Malaysia
(External Examiner)

Ajith Abraham, PhD
Professor
IT4Innovations- Center of Excellent VSB
Technical University of Ostrava
(External Examiner)

SITI HALIJAH SHARIFF, PhD
Associate Professor
Dean
Institute of Graduates Studies
Universiti Teknologi MARA
Date: 11th April 2016

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the result of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	: Raseeda Binti Hamzah
Student I.D. No.	: 2011803776
Programme	: Doctor of Philosophy of Science CS990
Faculty	: Computer and Mathematical Science
Thesis Title	: Discriminative Classification Model of Filled Pause and Elongation for Malay Language Spontaneous Speech
Signature of Student	:
Date	: April 2016

ABSTRACT

Automated speech recognition (ASR) for spontaneous speech poses extra challenge compared to read speech as it contains varied speaking rates, poor phonation and disfluencies. Studies have shown that filled pause is one of the most common disfluencies of spontaneous speech characteristic where it presents considerable problems for ASR performance. In many filled pause studies, the hindering factor is that filled pause being often recognized as short words which particularly has semantic meaning, such as ‘um’ can be recognized as ‘thumb’ or ‘arm’. This problem becomes especially pertinent where a vowel sound of normal word being relatively long at any position in an utterance, both within a word as well as between words which formerly known as elongation. The existence of elongation causes normal word falsely detected as filled pause due to their similar acoustical feature patterns. Classifying elongation as filled pause affects ASR’s performance as eliminating normal words from recognition may modify the intended context of a speech. Therefore, the main aim of this research is to classify filled pause and elongation into its own classes by constructing a discriminative classification model from the extracted acoustical features. A large number of signal features have been employed for the problem of discriminating filled pause and elongation. Several well-established features such as Formant Frequency (FF), Fundamental Frequency (F0), Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rates (ZCR) and Short Time Energy (STE) were used in this research. These features are carefully chosen to emphasize signal characteristics that differ between filled pause and elongation. In most speech research, extracting speech energy feature is still remains as challenging task due to it typically has a great deal of variance which include loudness as well as the variance in the signal energy between different phoneme which contains vowel or/and consonant sounds. One of the ways of detecting vowel and consonant is through its energy level. Beside the common way of quantifying the speech energy by calculating the sum of energy of the short interval centered on each interval, we proposed new technique namely, Local Maxima Energy (LM-E) to exploit the speech energy feature of filled pause and elongation. Experimentally, this can be done by measuring its amplitude transition from one frame to another by setting a threshold as height difference between peaks of the speech signal. Unlike other acoustical features, LM-E has shown its performance to classify elongation better by detecting the expressive contour of the elongation that is caused by the transition from consonant to vowel of the elongation. A rigorous feature performance evaluation shows that LM-E significantly increased the classification performance when fused with ZCR. Therefore, these two features are incorporated into discriminative Naïve-Bayes model for filled pause and elongation classification. The discriminative model of LM-E and ZCR improved the classification performance by 7% error rate reduction, and average of 7% accuracy increments compared to single feature classification performance. This model can further be used to improve disfluencies detection for a better ASR performance.

ACKNOWLEDGEMENT

In the name of Allah, the Most Beneficent, the Most Merciful. I am using this opportunity to express my gratitude to my beloved husband, family and friends who supported me throughout the journey of my PhD.

I really want to express my warm thanks to Prof. Madya Dr. Nursuriati Jamil and Dr. Noraini Seman for their full support and guidance at Universiti Teknologi MARA Shah Alam. I am thankful for their aspiring guidance, invaluable constructive criticism and friendly advice during the study. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the study.

Thank you also to the Universiti Teknologi MARA Shah Alam and Ministry of Education for supporting this study under the Tenaga Pengajar Muda UiTM and Skim Latihan Akademik Bumiputera scholarship.

I would also like to thank my virtual friends on Doctorate Support Group Malaysian postgraduate community.

This thesis will not be completed without all of these supports.

TABLE OF CONTENTS

	Page
CONFIRMATION BY PANEL OF EXAMINERS	ii
AUTHOR'S DECLARATION	iii
ABSTRACT	iv
ACKNOWLEDGEMENT	v
TABLE OF CONTENTS	vi
LIST OF TABLES	x
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xv
CHAPTER ONE: INTRODUCTION	1
1.1 Research Background	1
1.2 Problem Statement	2
1.3 Research Objectives	4
1.4 Research Scope	5
1.5 Research Contribution	5
1.6 Research Significance	6
1.7 Thesis Organization	7
CHAPTER TWO: LITERATURE REVIEW	9
2.1 Introduction	9
2.2 Spontaneous speech	10
2.2.1 Disfluencies	12
2.2.1.1 Repetition	14
2.2.1.2 Sentence Restart	14
2.2.1.3 Filled Pause	15
2.2.1.4 Elongations	16
2.3 Pre-Processing	18
2.3.1 Voice Activity Detection (VAD)	19
2.4 Speech Feature Extraction	20

2.4.1	Fundamental Frequency (F0)	23
2.4.2	Energy	24
2.4.3	Formant Frequency (FF)	24
2.4.4	Duration	25
2.4.5	Mel Frequency Cepstral Coefficients (MFCC)	26
2.4.6	Zero Crossing Rates (ZCR)	26
2.5	Pattern Classification	27
2.5.1	Naïve Bayes	32
	2.5.1.1 Bandwidth Parameter Selection for Kernel Density Estimation	37
2.6	Evaluation of Classification	39
2.7	Related Malay Language Spontaneous Speech Researches	41
2.8	Malay Language Speech Sounds and Rules	42
2.9	Summary	47
	CHAPTER THREE: METHODOLOGY	48
3.1	Introduction	48
3.2	Construction of Malay Filled Pause and Elongation Datasets	50
3.3	Malay Language Filled Pause	54
	3.3.1 Established Filled Pause Acoustical Characteristics	55
3.4	Malay Language Elongation	58
3.5	Pre-Processing	59
	3.5.1 Amplitude Normalization	59
	3.5.2 Pre-Emphasis	61
	3.5.3 Framing	62
	3.5.4 Windowing	63
	3.5.5 Voice Activity Detection	63
	3.5.5.1 Energy	64
	3.5.5.2 Zero Crossing Rates	67
	3.5.5.3 Energy and Higher-Order Differences	69
3.6	Summary	72

CHATER FOUR: FEATURE EXTRACTION FOR NEW ACOUSTICAL	
FEATURE CONSTRUCTION	74
4.1 Introduction	74
4.2 Standard Filled Pause Acoustical Features	74
4.2.1 Formant Frequency	75
4.2.2 Fundamental Frequency	77
4.2.3 Mel-Frequency Cepstral Coefficients	80
4.2.4 Zero Crossing Rates	82
4.2.5 Energy	82
4.3 Local Maxima of the Speech Energy	85
4.3.1 Local Maxima Extraction Using Minimum Peak Distance	
Threshold	86
4.3.2 Local Maxima Extraction Using Minimum Peak Height Threshold	88
4.3.3 Proposed Local Maxima Threshold Selection	89
4.4 Feature Ranking	95
4.5 Summary	97
CHAPTER FIVE: PATTERN CLASSIFICATION FOR CONSTRUCTION OF	
DISCRIMINATIVE MODEL	98
5.1 Introduction	98
5.2 Bayes Classification Process	98
5.2.1 Conditional Probability Density Estimation	100
5.2.1.1 Kernel Density Function Using Silverman’s Bandwidth	101
5.3 Acoustical Rules Parameter Selection for Filled Pause and Elongation	102
5.4 Cross Validation	104
5.6 Discriminative Classification Model Construction Using Naïve Bayes	106
5.7 Evaluation Method	109
5.8 Summary	110
CHAPTER SIX: RESULTS AND DISCUSSION	111
6.1 Introduction	111
6.2 Voice Activity Detection Results for Noise Removal	111
6.3 Feature Ranking for Pause and Elongation Representation	116

6.4	Single Feature Bayes Classification Performance Evaluation	117
6.5	Discriminative Classification Performance Evaluation	127
6.6	Classification Performance Comparison with English Language Datasets	134
6.7	Summary	137
CHAPTER SEVEN: CONCLUSION AND FUTURE WORKS		138
7.1	Introduction	138
7.2	Review of Objectives	138
7.3	Summary of Research Findings	139
7.4	Significant Contribution of the Research	140
7.5	Research Limitation	141
7.6	Future Research Enhancement	142
REFERENCES		144
APPENDICES		160
AUTHOR'S PROFILE		165

LIST OF TABLES

Tables	Title	Page
Table 2.1	Type of Spontaneous Speech Recording Used In Spontaneous Speech Studies	11
Table 2.2	Example of Disfluencies	14
Table 2.3	Filled Pause Example Uttered On Language Basis	15
Table 2.4	Acoustical Rule-Based Filled Pause Researches	22
Table 2.5	Filled Pause Research Based On Acoustical and Classification Method	30
Table 2.6	Confusion Matrix of Classifier Evaluation	40
Table 2.7	Examples of Malay Word Alteration and the Transformation	43
Table 3.1	Quantitative Information of Selected MPHD Files	52
Table 3.2	Average Standard Deviations of Acoustical Features Measurements of Filled Pause and Normal Words	56
Table 3.3	Common Elongated Word in the MPHD Data	59
Table 3.4	Mean Amplitude Variance Due To Normalization	61
Table 3.5	Statistical Information of the HOD Method of VAD	69
Table 3.6	Disfluencies Datasets for Filled Pause and Elongation	72
Table 4.1	Standard Deviation Comparison between LM-E Techniques	91
Table 4.2	The Acoustical Features and Its Notations	96
Table 5.1	Proposed Acoustical Rules for Filled Pause and Elongation Classification	103
Table 6.1	Effects on Formant Frequencies Standard Deviation before and After Unwanted Speech Interval Removal for Energy-Based VAD	113
Table 6.2	Effects on Formant Frequencies Standard Deviation before and After Unvoiced Removal for Energy and ZCR Based-VAD	114
Table 6.3	Formant Frequencies Standard Deviation Comparison	115
Table 6.4	Mean of Formant Frequencies Standard Deviation of Each Datasets	116
Table 6.5	The Z-Score for Each Acoustical Features and Decision about It's Importance	117

Table 6.6	F-Measure of 10 Fold CV on Single Feature Classification	120
Table 6.7	Precision of 10 Fold CV on Single Feature Classification	120
Table 6.8	Recall of 10 Fold CV on Single Feature Classification	121
Table 6.9	Accuracy of 10-Folds Cross Validation on the Single Feature Classification	121
Table 6.10	10-Folds CV Accuracy (%) Of Discriminative Classification Using Combined Data	129
Table 6.11	F-Measure for 10 Fold CV for Combined Data	130
Table 6.12	Precision of the 10-Fold CV for Combined Data	131
Table 6.13	Recall Rate of 10-Fold CV for Combined Data	132
Table 6.14	Information of Probability Density Estimation of FP11.Wav and ELO07.Wav	133
Table 6.15	English Elongations	134
Table 6.16	Discriminative Model (LM-E+ZCR) Classification Performance on English Data	136
Table 6.17	Comparison of Discriminative Model (LM-E+ZCR) Classification on Malay and English Language Datasets	137

LIST OF FIGURES

Figures	Title	Page
Figure 2.1	Overview of Literature Review Topic Coverage	9
Figure 2.2	Basic Disfluencies Framework	10
Figure 2.3	Example of Disfluencies in Spontaneous Speech	13
Figure 2.4	Kernel Density Estimation Example (Silverman, 1986)	37
Figure 2.5	KDE Plot of Various Bandwidth Methods	39
Figure 2.6	Structure of Malay Language Phonemes	44
Figure 2.7	Human Vocal Mechanism	45
Figure 3.1	Basic Flow of Research Methodology	48
Figure 3.2	Research Framework	49
Figure 3.3	Data Collection Process Flow	51
Figure 3.4	A Complete Sentence with Occurrences of Filled Pause Only (Malay Sentence Id S169M9T04: Pesakit <i>aaa</i> , Buah Pinggang)	53
Figure 3.5	A Complete Sentence with Occurrence of Filled Pause, Normal Words and Elongation (Malay Sentence Id S53M5T03: Di (ELO) <i>aaa</i> (FP) Negara Jiran)	53
Figure 3.6	Filled Pause <i>/aaa.Wav/</i>	54
Figure 3.7	Filled Pause <i>/eee.Wav/</i>	54
Figure 3.8	Filled Pause <i>/emm.Wav/</i>	55
Figure 3.9	Robust Filled Pause	57
Figure 3.10	Non Robust Filled Pause with Short Duration and Having Word Insertion	57
Figure 3.11	Non Robust Filled Pause with Expressive Intonation Contour	57
Figure 3.12	Elongation Word at the Last Syllable of <i>/Da/</i> From Word <i>/A//Da/</i>	58
Figure 3.13	Steps of Amplitude Normalization on the Speech Signal	60
Figure 3.14	Location of Upper and Lower Energy Threshold	67
Figure 3.15	Location of N1 and N2 Level	69
Figure 3.16	Effect of Different Order of HOD Algorithm on VAD	71
Figure 4.1	Acoustical Feature Extraction Process for Each Speech Segment	74

Figure 4.2	Example of Formant Frequency Plot of Elongation of the Word /Di/	77
Figure 4.3	Process of F0 Detection by Using Autocorrelation Function	79
Figure 4.4	Mel-Frequency Cepstral Coefficients Extraction Process	80
Figure 4.5	Example of STE Measurements on Elongation	83
Figure 4.6	Example of STE Measurements on Filled Pause	84
Figure 4.7	STE Value Distribution for Filled Pause and Elongation	85
Figure 4.8	Local Maxima Location	86
Figure 4.9	Example of Local Maxima Extraction by Using Minimum Peak Distance as Threshold	87
Figure 4.10	Distribution of Local Maxima Using Minimum Peak Distance	88
Figure 4.11	Distribution of Local Maxima Using Minimum Peak Height	89
Figure 4.12	Example of Proposed Local Maxima Extraction	90
Figure 4.13	Filled Pause (A) And Elongation (B) From 28082008 Dataset	91
Figure 4.14	Proposed LM-E Data Distribution of Filled Pause and Elongation	92
Figure 4.15	Flowchart of the Proposed LM-E Technique	94
Figure 4.16	Example of LM-E Measurements on Filled Pause and Elongation	95
Figure 5.1	Single Classification Process	99
Figure 5.2	ZCR Histogram for Filled Pauses	100
Figure 5.3	ZCR Histogram for Elongations	100
Figure 5.4	ZCR Kernel Density Plot for Filled Pauses and Elongations	101
Figure 5.5	KDE of LM-E for Filled Pauses and Elongations	103
Figure 5.6	Acoustical Rules Pseudo Code	104
Figure 5.7	Ten-Fold Cross-Validation Process	106
Figure 5.8	Naïve Bayes Illustration	107
Figure 5.9	Illustration of Naïve Bayes Fusion Classification	108
Figure 6.1	Voice Region Detected By Energy-Based VAD	112
Figure 6.2	Voice Region Detected By Energy + ZCR-Based VAD	113
Figure 6.3	Voice Region Detected By Energy-HOD Based VAD	115
Figure 6.4	F-Measure of 10-Fold CV	118
Figure 6.5	Recall and Precision Average of 10-Fold CV	119
Figure 6.6	Average Accuracy of Each Acoustical Feature	122
Figure 6.7	Example of Misclassified Filled Pause	123

Figure 6.8	Example of Misclassified Elongation	124
Figure 6.9	Consonant to Vowel Transition in Elongation /Da/	125
Figure 6.10	Kernel Density Estimation of STE for Filled Pause and Elongation	126
Figure 6.11	English Language Elongation of the Word ‘That’	135
Figure 6.12	English Language Elongation of the Word ‘Showers’	135
Figure 6.13	English Language Elongation of the Syllable ‘Ver’	136

LIST OF ABBREVIATIONS

Abbreviations

ELO	Elongation
FF	Formant Frequency
F0	Fundamental Frequency
FP	Filled pause
HOD	Higher Order Differences
KDE	Kernel Density Estimation
LM-E	Local Maxima of the Speech Energy
MFCC	Mel Frequency Cepstral Coefficients
MPHD	Malaysian Parliamentary Hansard Document
MSD	Manual Sentence Data
MZSA	Maximum Z-score among Shadow Attribute
NB	Naïve Bayes
STE	Short Time Energy
VAD	Voice Activity Detection
ZCR	Zero Crossing Rates

CHAPTER ONE

INTRODUCTION

1.1 RESEARCH BACKGROUND

The effort to recognize speech using technology started as early as the second half of the 20th century (Rabiner, 1977). Automatic speech recognition (ASR) attracted a lot of interest commercially since the first effort to achieve it started in the 1900s. Recently, the inclusion of speech technologies into applications to improve human productivity has noticeably changed life habits of many people. Lately, almost everyone is equipped with electronic gadgets, such as PDAs, smart phone and personal computers that have the computational technology necessary for speech recognition and reproduction. Most of the applications installed in each of the gadgets have the ability to recognize and execute voice commands. This machine ability is called as speech to text process that is the main task of ASR.

Despite recent advances, the ability of ASR to process speech is still well behind human capability. ASR needs significant improvement, especially in the processing of natural or colloquial speech, to reach levels of accuracy close to human capability. There are several excellent software and hardware combinations, such as IBM via voice, Dragon Naturally Speaking, and Philip's Free Speech available and used for human task simplification. Nevertheless, these tools are still behind expectations because of the absence of natural spoken processing ability. A crucial goal of ASR is to establish the ability to process natural spoken language in natural environments. In order to achieve this goal, many obstacles and challenges need to be resolved, particularly regarding natural speech.

Natural speech, also known as spontaneous speech, is uttered freely without any plan or rules. This is the kind of speech with which people are normally familiar and use without thinking about it. The processing of this kind of spontaneous speech needs extra processing expertise because it involves the human auditory system inclusive of linguistic, paralinguistic and extra-linguistic components. Unlike purposeful speech events like reading out loud, spontaneous speech is produced at

faster speaking rates, poor pronunciation, contains frequent grammatical errors, and can contain many disfluencies. These characteristics increase the challenge of ASR computation and reduce its utility. Studies showed that the disfluencies that occur in spontaneous speech is one of the obstacles that ASR needs to solve if it is to further progress in functionality and further integrate into people's daily work (Kaushik et al., 2010). It is believed that removing or detecting disfluencies such as filled pause can improve ASR (Stouten, 2008).

1.2 PROBLEM STATEMENT

Speech recognition degrades with the occurrences of the filled pause because it interrupts the fluency of speech, increases ASR complexity, and causes confusion to machine based recognition devices (Stouten, 2008). From the speech recognizer's point of view, there are two ways of handling a filled pause. The first is to remove the filled pauses to ensure the smoothness of the recognizer's processing. The second is to model the filled pause and allocate it to part of the decoding process. The first approach is much harder to deal with and cause considerable problems for speech recognition since filled pause is often being recognized as short words such as an "um" can be mistaken as "thumb" or "arm". This problem becomes pertinent when a vowel sound of a normal word being spoken relatively long at any position in an utterance, both within a word as well as between words. This occurrence formerly known as elongation causes a normal word to be falsely detected as filled pause because both elongation and filled pause shared similar acoustical feature patterns (Kaushik et al., 2010). Several established related researches have been conducted in detecting the filled pause, where both filled pause and elongation were classified into the same disfluency class (Moniz, et al., 2013; Audhkhasi & Angeles 2009; Stouten & Martens, 2003; Goto et al., 1999). However, classifying filled pause and elongation into the same disfluency class can affect ASR's performance as eliminating normal words from recognition may modify the intended context of a speech and leads to inaccurate transcription. According to Kaushik et al., (2010), filled pause and elongation causes transcription problem in ASR.

Generally, filled pause is non-lexical speech, which means it does not have any vocabulary meaning. Alternatively, elongation does have its own denotation.

Therefore, classifying elongation as a kind of filled pause may cause semantic change in a sentence. For example, in the Malay sentence “*Saya akan bertanggungjawab terhadap kes tersebut*” translated in English means “I will be responsible for that case”. In this sentence, the second syllable /ya/ of the word is prolonged as elongation. If a Malay language ASR detected and removed the elongated syllable in the word /saya/, the sentence will become “*sa akan bertanggungjawab terhadap kes tersebut*” in English “(/sa/ no_meaning) will be responsible for that case”. Therefore, the first word ‘sa’ has no meaning and the entire sentence semantic is changed. Thus, this research emphasized on classifying elongation of Malay spontaneous speech as normal words, while the filled pause is classified as disfluency.

Further work in classifying filled pause and elongation as separate classes has received interest lately, as found in (Li et al., 2008, 2010; Veiga, 2011; Verkhodanova & Shapranov, 2014). Li et al. (2008) used Fundamental Frequency (F0), Short Time Energy (STE), and Mel Frequency Cepstral Coefficient (MFCC) to classify filled pause and elongation using Hidden Markov Model classifier. Veiga (2011), on the other hand, utilized F0, energy, duration, and spectral envelope for Portuguese language filled pause and elongation classification.

Among the well-established acoustical features, F0 is mostly used as can be found in Gabrea et al., (1999); Goto, et al., (1999); Gaurav & Nigel, (2006); Audhkhasi & Angeles, (2009); Ogata, Goto, & Itou, (2009); Kaushik et al., (2010); Karpiński, (2013); Verkhodanova & Shapranov, (2014). F0 is associated with energy as confirmed by Rosenberg & Hirschberg, (2006) in his work where energy is used to classify pitch into accented or non-accented word. Energy of the speech may be measured using several techniques such as log energy, sum of square energy and sum of absolute energy. Generally, all the above-mentioned techniques of calculating the sums of energy are measured on each short frame. These techniques are suitable and beneficial for speech involving normal words. However, sum of energy cannot sufficiently represent filled pause, especially when filled pause needs to be differentiated with elongation. According to Veiga (2011), the current means of representing energy is not able to separate filled pause and elongation in Portuguese language well due to their similar energy characteristics. The use of energy parameter is not limited in endpoint detection only. It is also beneficial in consonant and vowel detection in Izzad et al., (2013); Mercier et al., (1989). However, sum of energy

calculated from short time speech frame is unable to detect the energy variation from the consonant and vowel in the elongation. These researchers concluded that there are difficulties in differentiating filled pause and elongation into two separate classes. Therefore, further work is needed to investigate and select the suitable acoustical features for the abovementioned purpose. Rigorous acoustical feature selection research for representing filled pause and elongation remains hard to find. Therefore, this research aims to identify the acoustical feature of filled pause and elongation, and construct a classification model that is able to discriminate filled pause and elongation into their own separate classes.

Generally, the problem statements of this research can be summarized as below:

- i) The Malay filled pause and elongation datasets are hardly available.
- ii) The use of existing speech energy extraction technique is unable to differentiate filled pause and elongation.
- iii) The existing acoustical model of filled pause and elongation is unable to discriminate filled pause and elongation into two separate classes.

1.3 RESEARCH OBJECTIVES

Based upon the abovementioned problems, the main aim of this research is to classify filled pause and elongation disfluencies of Malay language spontaneous speech into separate classes. In order to achieve this primary outcome, several specific objectives are executed as follows:

- i) **Objective 1:** To create Malay language's filled pause and elongation dataset from spontaneous speech.
- ii) **Objective 2:** To produce a new energy feature extraction technique for filled pause and elongation.
- iii) **Objective 3:** To model the discriminative properties of filled pause and elongation acoustical features

1.4 RESEARCH SCOPE

This research focuses on filled pause because previous research and preliminary analysis showed that filled pause is the highest occurred disfluency in any spontaneous speech and its existence degrades the ASR accuracy. Elongation is also another disfluency included as it is most confused with filled pause and is the main motivation of this research. The speech data collection that is used in this research is the Malaysian Parliamentary Hansard Document (MPHD) sessions for the year 2008 (Seman, 2008). The MPHD data consists of 22 sessions totaling to 198 hours of parliament debate sessions of Dewan Rakyat. Each session is represented in a huge raw video file and manually transcribed text. The MPHD video and audio contain many different types of noises and non-speech elements, such as filled pause, laughter, cough, claps and also noises from the microphone. In this research, the focus is on the audio wave signal that is extracted manually from the video. The Malay language that is used in the MPHD is a standard Malay spontaneous speech interlaced with some English language and used with permission from the Deputy of the Parliament session. However, for this research, only the Malay spoken speech in the overall 198 (22 sessions X 9 hours) session is utilized. The MPHD also contains a read speech session, which is known as the introduction session at the beginning of the debate. This read speech session is not included in this research because the focus is on spontaneous speech only.

A total of 1348 spoken sentences are gathered, with each sentence comprising at least three filled pauses and two elongations. Three types of filled pause pronunciations variations that are ‘aaa’, ‘eee’, and ‘emm’ are covered. Each type of filled pause variations is uttered by different speakers. The final data set consists of 3000 filled pauses and 3000 elongations.

1.5 RESEARCH CONTRIBUTION

Several contributions are derived from this research. The main contribution is the construction of a discriminative classification model of Malay filled pause and elongation. As stated by Vlasenko (2012), the linguistic content of an utterance is important to improve spontaneous ASR. However, disfluencies and additional

information of the utterance by the speakers cannot be neglected in order to improve spontaneous speech recognition.

The contributions of this research are listed as:

- i. The first contribution is the creation of the Malay filled pause and elongation datasets. Malay language filled pause and elongation datasets are collected manually and referred to as Malay filled pause dataset (FP_DATA) and Malay elongation dataset (ELO_DATA). From the manual data, three types of voice activity detection (VAD) are applied and compared to get an exact representation of filled pause and elongation segments.
- ii. A local maxima of the speech energy (LM-E) acoustical feature to represent the expressive contour of the elongation is introduced. LM-E is incorporated as one of the contributing features in the discriminative classification model and showed significant performance improvement of filled pause classification by increasing the accuracy by 7%. The improved LM-E has also enabled the classifier to differentiate elongation better with an improved average accuracy of 81%.
- iii. A set of acoustical feature-based rules are further developed by the estimated thresholds from Kernel density estimation. Then, a new discriminative classification model is composed based on the developed rules. The developed Kernel-based acoustical features rules are presented in *Section 6.3*, Chapter Six.

1.6 RESEARCH SIGNIFICANCE

The significance of this research can be observed from different perspectives. The first significance is primarily for automatic speech recognition enhancements. As proven in many research, filled pause is one of the main problems in spontaneous speech recognition. The detection of filled pauses should reduce the recognition error by removing it prior to speech recognition processing. Filled pause detection is meant for detecting a filled pause in a sequence of spontaneous speech sentences constructed with normal words, silences and other disfluencies. In order to detect a filled pause, a correct classification of filled pause and elongation is needed so that the detection will

become more accurate. It can also be inserted in the dictionary of the speech recognizer to be recognized in the same manner of the normal words.

Besides improving ASR's performance, filled pause detection can also be found in other application such as human mental state analysis (Gaurav & Nigel, 2006), fraud analysis (Humphreys, 2010) and discourse analysis (Karpinski, 2013). The filled pause detection can be used in other areas such as psychology. Psychology is a discipline that has interest in perception, cognition, attention, emotion, intelligence, phenomenology, motivation, brain functioning, personality, behavior, and interpersonal relationships, including psychological resilience, family resilience, and other areas. It is stated that filled pause indicates uncertainty of the speaker. In (Humpherys, 2010), fraud detection research has been done by analyzing the occurrence of filled pauses in speech. As such, the occurrence of filled pause is one of the cues of deceptive communication (Humpherys, 2010). Another application of filled pause research is in speech fluency analysis (Gaurav & Nigel, 2006). Filled pause shows a speaker's lack of knowledge on the discussed topic and may be caused by a lack of preparation (Audhkhasi & Angeles, 2009). According to Clark & Fox, (2002), filled pauses are the interpretation of several elements of the speakers mental state, which are:

- i. Thinking of the previous words uttered
- ii. Inviting others to think about the uttered words
- iii. Showing politeness
- iv. Completed, ceded, or maintaining the conversation
- v. Needing help to complete the conversation
- vi. Inviting others present to speak
- vii. Asking for a turn to speak
- viii. Marking discourse boundaries
- ix. Providing information

1.7 THESIS ORGANIZATION

This thesis presents a wide spectrum of work in detecting spontaneous filled pauses and elongation in spontaneous speech. Chapter One summarizes the whole research starting from the background, motivation, aim, scope, contribution,

significance of the research towards the improvement of existing filled pause and elongation classification. Chapter Two discusses related research inclusive of the detailed discussion of filled pause signals such as the types, the language basis of the filled pause, the pattern, and its usage. It also covers the techniques involved in filled pause and elongation researches. Chapter Three covers mainly the research methodology. In this chapter, the preprocessing of the data, which is the starting of the entire methodology, is presented. The thesis then continues with Chapter Four which presents various acoustical features extraction of filled pause and elongation. Nine acoustical features are discussed in this chapter. Then, Chapter Five discusses the classification technique consisting of two types of classification which is single pattern and multiple pattern classification. The results and discussion of each stage of methodology is then presented in Chapter Six. The thesis is then concluded in the last chapter which is Chapter Seven.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

Automatic speech recognition (ASR) is a process of converting speech signal to text. Read speech signals are easier to process compared to speech surrounded with disfluencies (spontaneous speech). Filled pause is one of the most occurred disfluencies that need to be detected to avoid complication during ASR. This chapter discussed the literature review of this research that is basically divided into five sections. The overall literature review of this thesis is described as in Figure 2.1.

The first section is the review on spontaneous speech. This section aims to discuss the related topics of spontaneous speech such as spontaneous data collection, types of disfluencies, previous filled pause researches and the Malay language spontaneous speech researches. The second section discussed the pre-processing in automatic speech recognition (ASR) system. The third section presents the feature extraction techniques related to filled pause researches. Furthermore, the reviews of classifications are discussed in the fourth section. The fifth section presented the previous evaluation methods in filled pause research.

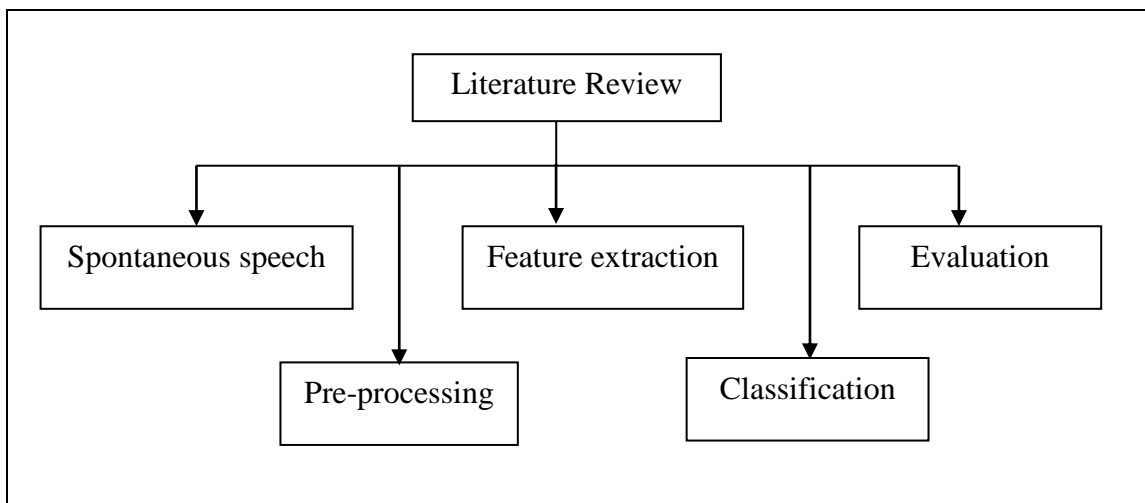


Figure 2.1: Overview of literature review topic coverage

2.2 SPONTANEOUS SPEECH

Spontaneous speech is speech that is uttered with no available text prepared beforehand and the speakers speak freely without any rules. In comparison with read speech, it is planned and prepared texts uttered by speakers making it less error and improve fluent speech. Spontaneous speech is different from read or rehearsed speech partly because of the disfluencies produced by a speaker during spoken language production (Yildirim, 2009). Figure 2.2 shows the division of human speech that is categorized into read and spontaneous speech.

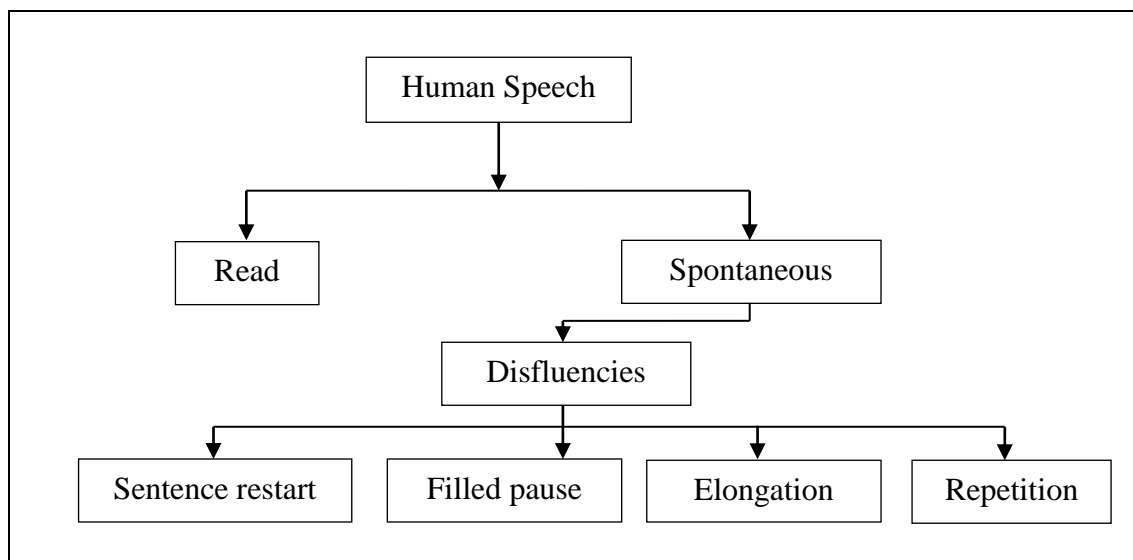


Figure 2.2: Basic disfluencies framework

Processing spontaneous speech is much harder compared to read speech as it contains disfluencies (Ogata et al., 2009; Stouten et al., 2006). The occurrences of disfluencies disturb the normal sequence of the sentence and disrupt the performance of ASR. Spontaneous speech can occur in any forms of conversation. Previously, several disfluencies researches have used different types of corpuses. The corpuses that were used in spontaneous speech studies are illustrated in Table 2.1.

A number of researchers has developed and used spontaneous speech corpus from different types of conversation such as lectures, presentation, and new commentaries (Furui et al., 2005). Another corpus was constructed by the same group (Nakamura et al., 2008) with different raw materials such as academic presentations, extemporaneous presentations, and dialogues. They found that, the reduced distance

between phonemes is the key measurement to differentiate spontaneous with read speech, thus concluded that spontaneous speech is acoustically different from read speech.

Table 2.1:

Type of spontaneous speech recording used in spontaneous speech studies

Authors (Year)	Language	Type of spontaneous speech recordings
Goto et al., 1999	Japanese	Japanese spontaneous speech corpus
Yokoyama et al., 2003	Japanese	Different speakers and topic presentation
Byrne, 2004	English Russian Czech	Interview with specific purpose that covers basic topic of human history.
Furui, 2005	Japanese	Monologues such as lectures, presentations and news commentaries.
Nunes & Neves, 2006	Portuguese	A corpus of oral interviews collected by the Center of Linguistics of the University of Lisbon
Nakamura et al., 2007	Japanese	Japanese spontaneous speech includes academic presentations, extemporaneous presentations, and dialogues.
Li et al., 2008	Mandarin	Chinese Annotated Dialogue and Conversation Corpus
Audkhasi & Angeles, 2009	English	Life recordings of 96 candidates who were interviewed for call centre agent positions at IBM Daksh's Gurgaon India call centre facility.
Seman et al., 2010	Malay	Malaysian Parliament Debate Database
Veiga, 2011	Portuguese	Podcasted television news, resulting in around 22 hours of non-annotated speech
Chong et al., 2012	Malay	Conversational speech with prepared topic to different speakers
Karpiński, 2013	Polish	"Origami" dialogue session recordings
Verkhodanova & Shapranov, 2014	Russian	Map-tasks and appointment-task dialog

Honal & Schultz, (2005) defined spontaneous speech as unsmooth speech. They constructed a spontaneous corpus from scheduled, fictitious meeting between two persons. The same phenomenon such as filled pause, sentence restart, and repetition are collected from Slovenian news to improve the spontaneous ASR performance (Zgank & Maucec, 2010). Yokoyama et al., (2003) used a presentation of 10 male speakers as a spontaneous Japanese corpus in their research. Spontaneous presentation utterances are both acoustically and linguistically variable according to speakers and topics.

Byrne 2004 described spontaneous speech by conducting an interview with specific purpose that covers basic topic of human history. There is a structure to the interviews which begin with education, occupation, living conditions, life in the campus and liberation. It is still in the scope of spontaneous speech even though the speaker really know what they should converse in general. This is because their conversation produces disfluencies which show the characteristics of spontaneous speech.

In Chong et al., (2012) a Malay spontaneous corpus was developed from a given topic to different speakers which include normal words, code switched word, proper nouns and disfluencies. Similar to other languages, filled pause is the most occurred disfluency in spontaneous speech. In their research, about 11 000 filled pauses along with multiple filler words were observed. However, no further focus processing is taken on the collected filled pauses except manual transcription.

A Malaysian Parliament Hansard Data (MPHD) for spontaneous speech recognition of the year 2008 was used by Seman et al., (2010). In the database, there are two types of speeches which consists of read and spontaneous speech. In their research, only the spontaneous speech is considered. The spontaneous speech component comprises debate speech session that produces a lot of disfluencies such as filled pause, elongation, repetition and sentence restart. The issues or topic that is being debated in the Malaysian parliament is prepared. However, the content delivery is spontaneously discussed. The same database (i.e MPHD) is used in this research, which focus on the filled pauses and elongations occurrences.

2.2.1 Disfluencies

Disfluency is a common phenomenon for free-rules speech or spontaneous speech (Lin & Lee, 2009). Figure 2.3 shows an example of a Malay language spontaneous sentence containing disfluencies.

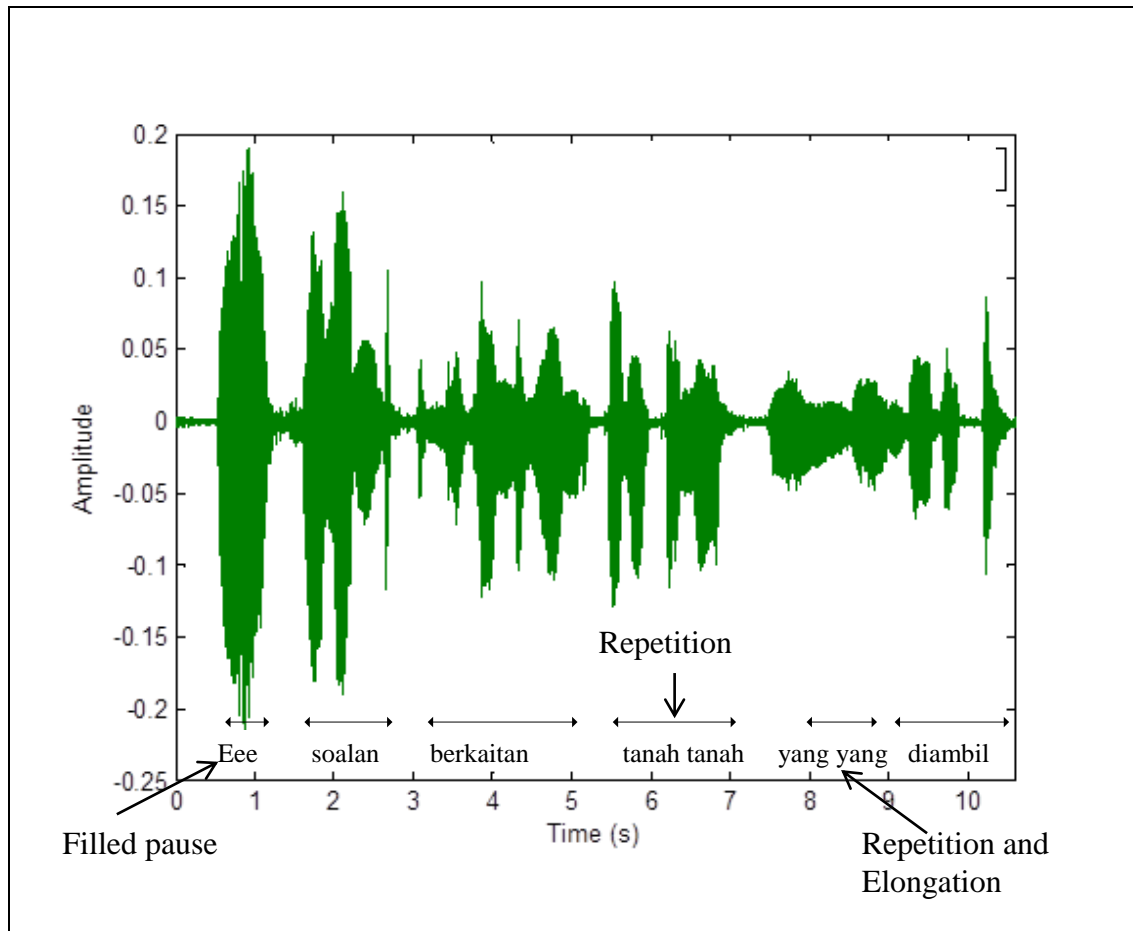


Figure 2.3: Example of disfluencies in spontaneous speech

The filled pause is seen in *eee*, while repetition is *tanah tanah* and *yang yang*. They can be classified as repetition, sentence restart and filled pause. However, there are differences in terminology used for each disfluencies (Shriberg, 1994). For example, filler for filled pause is also denoted as insertion, interjection and hesitation (Kitagawa et al., 2011); repairs for sentence restart (Heeman & Allen, 1999); repeated word or repetition (Shriberg et al., 1997) and prolongation or lengthening for elongation (Eklund, 2001). Among them, filled pause is the most common types of disfluency and has attracted attention of disfluencies researches. The corpuses that were used in disfluencies research are mostly conversational speeches such as meeting, lectures and dialogs which were spoken freely without any prepared text. Since spontaneous speech is spoken freely, speakers need time to think the next word to be spoken. The gap of thinking is usually filled with filled pause. Table 2.2 lists few examples of disfluencies. Further discussion of each type of disfluencies follows in the next subsection.

Table 2.2:
Example of disfluencies

Type of Disfluencies	Example
Filled Pause (Zgank & Maucec, 2010)	‘Um’, ‘eee’, ‘uh’
Repetition (Stolcke & Shriberg, 1996)	‘It’s a it’s a fairly large community’ -Repeated word: <i>it’s</i>
Sentence Restart (Stouten et al., 2006)	‘ <i>In a situation with uh</i> ’ in a country with two speed levels’
Elongation (Eklund, 2001)	“Book”, “hotel”, and “taxi”

Disfluency in speech is non-lexical vocables or any form of breaks, irregularities that occur in spontaneous speech. It is occasionally used in an instance to rephrase spoken sentences or to correct mispronunciation during spontaneous speeches. Disfluencies are categorised as filled pause, repetition, sentence restart, silent pauses, hesitations, elongation, truncations, and self-repairs. Among them, filled pause is the most common (Chong et al., 2012).

2.2.1.1 Repetition

One of the strategies for a speaker to gain some time to think is to repeat a word once or several times before continuing with the rest of the sentence. If this happens, the last word of the repeated word sequence is considered as the regular word whereas the others are designated as disfluencies. However, in Malay language, there is a reduplicative word type such as *sama-sama* in English (you are welcome), *mata-mata* (policeman), and *kanak-kanak* (kids) (Semana et al., 2010). The reduplicative word can be original word or to show multiple things.

2.2.1.2 Sentence Restart

Sentence restart is a situation in which the speaker aborts the current utterances and start over with a new sentence in order to make a correct statement. An example of sentence restart is shown in the following sentence spoken by a speaker B. Speaker B: *aaa (filled pause)* sebenarnya kita. ***Kami*** di Jabatan Perhutanan. The word *Kami* printed in bold italic indicates the sentence restart. From the example, the speaker stops at the word ‘*kita*’ and restart the sentence to specifically mention who is meant for ‘*kita*’.

2.2.1.3 Filled Pause

Generally, filled pause is a verbal device that has no lexical meaning such as ‘um’ and ‘uh’ (Gaurav & Nigel, 2006), while normal words is a combination of consonant and vowels for example /work/, /yes/, and /no/. The existence of filled pauses in spontaneous speech has been studied from as early as 1992 (O’Shaughnessy, 1992). Compared to repetition and sentence restart, filled pause is one of the disfluencies recognized as the highest existence in spontaneous speech (Medeiros, Moniz, et al., 2013). Filled pause is usually used by speakers to prevent interruption from others while planning their utterance (Zuraida & Knowles 2006; Goto et al., 1999). In the area of speech recognition, researchers have found that filled pause is one of the factors that degrades ASR (Li et al., 2008, Peters, 2006; Kaushik et al., 2010). In language studies, filled pause indicates speakers state of thinking (Gaurav & Nigel, 2006) and greater rhetorical and emphatic qualities of the spoken utterances (O’Connell & Kowal, 2004). Filled pause is also being used to evaluate prosody and turn-taking in conversation (Begum et al., 2008; Zuraidah & Knowles, 2006).

Filled pause is uttered based on language. An English-standard filled pause is described as /uh/ and /um/ (O’Shaughnessy, 1992), while for other languages as found in Dutch are /De/, /Die/ (Stouten & Martens, 2003), Mandarin /e/, /en/ (Li et al., 2008) and Portuguese /uum/, /aaa/, /eee/ (Veiga et al., 2011). The example of filled pause based on language is displayed in Table 2.3.

Table 2.3:
Filled pause example uttered on language basis

Language	Filled pause
English	‘Uh’, ‘Um’
Dutch	‘De’, ‘Die’
Portuguese	‘Uum’, ‘aaa’, ‘eee’
Japanese	‘Ee’, ‘maa’, ‘ano’
Mandarin	‘Ah’, ‘ung’, ‘um’, ‘em’, ‘hem’
Slovenian	‘eee’, ‘mhm’, ‘aaa’, ‘sss’, ‘mmm’
Malay	‘ah’, ‘eh’, ‘ur’, ‘um’, ‘mhm’
Russian	‘e’, ‘ə’, ‘m’, ‘vot’, ‘nu’

In English and other languages, filled pause has been widely studied by the research communities (Veiga et al., 2011; Wu & Yan 2004; Lee et al., 2004). However, for Malay Language, very few language studies are focused on filled pause (Begum et al., 2008; Zuraida and Knowles, 2006). Although the work done in

Zuraidah & Knowles, (2006) is not mainly focus on filled pause phenomena, they have shown the importance of filled pause in spontaneous speech as indication of possible turn-taking in conversation. Therefore, in this research the filled pause is chosen due to the high occurrences and its importance in spontaneous speech.

Filled pause is written in many forms depending on the language such as /emm/, /aaa/ and /eee/ for the Malay language. Filled pauses in the English language are described as /uh/ and /um/ (Gabrea et al., 1999). In other languages, filled pauses are /De/ and /Die/ in Dutch (Stouten & Martens 2003), /e/ and/en/ in Mandarin (Li et al., 2008) and for Portuguese /uum/, /aaa/ and /eee/ (Veiga et al., 2011). Although the filled pause is language dependent, their acoustical features patterns are the same (Audhkhasi & Angeles, 2009; Kaushik et al., 2010). The most common acoustical feature characteristics are flat fundamental frequency, stable formant frequencies, constant energy, longer duration, adjacent silence, spectral center of gravity, stable interval durations, lengthening property, and nasal property (Wu & Yan, 2004).

There were various languages involved in filled pause researches and were done for many purposes in many fields. In fraud investigation (Humpherys, 2010), filled pause is one of the speech event detected to analyze fraud based on the theory that deceptive speech is cognitively more problematic compared to honest speech (Buller & Burgoon, 1996; Ekman & O'Sullivan, 1991). In the field of psychology, filled pause counts are used to indicate the speaker's fluency stage either in the aspects of language usage or knowledge on the topic (Gaurav & Nigel, 2006; Moniz, et al., 2009) (Gaurav & Nigel, 2006; Moniz, Trancoso, & Mata, 2009). In language studies, filled pause is an indicator that shows speakers state of thinking (Gaurav & Nigel, 2006) and greater rhetorical and emphatic qualities of the spoken utterances (O'Connell & Kowal, 2005). Filled pause is also being used to evaluate prosody and turn-taking in conversation (Zuraidah & Knowles, 2006).

2.2.1.4 Elongations

Another disfluency which is known as elongation is considered similar to filled pause (Deme & Marco, 2013). An elongation is defined as a word that has phonetic prolongation, that is a word that contains a lengthened vowel in the word (Ogata et al., 2009). For example, in the Portuguese language, the phonetic /v/ in the

word /para/ is uttered longer in duration than usual (Veiga et al., 2011). In English, elongated vowels are found at the start and the end of some words such as /Lomaland/, /Paraguay/, /feel/ and /field/ (Gaurav & Nigel, 2006; Kaushik et al., 2010). In Mandarin, the following words are observed to be elongated by speakers ‘个’ (/ge/), ‘是’ (/shi/), ‘的’ (/de/) and ‘在’ (/zai/) (Li et al., 2008). In the Malay language, the vowel /e/ at the end of the word such as /saya/, /ada/, and /nya/ is always elongated by speakers. Similarly, filled pause and elongation are also being used by the speakers as a tool to maintain their turn in conversation.

Elongation is defined as overextended speech segments that can be found in a word or syllable (Eklund, 2001). In most disfluencies detection researches, elongation is categorized as filled pause (Verkhodanova & Shapranov, 2014; Zgank & Maucec, 2010; Audhkhasi & Angeles, 2009; Gaurav & Nigel, 2006; Lee et al., 2004; Eklund, 2001; Goto et al., 1999). Based on Eklund (2001), elongation is language dependent and has the same characteristics as filled pause in term of its duration and vocalization. Elongations serve the same purpose of filled pause. In spontaneous speech, speakers will either use filled pause or elongation to fill the gap of thinking. Although elongation serves the same function as filled pause, elongation has its own semantic meaning while filled pause is meaningless.

The occurrences of filled pause and elongations disturb the accuracy of ASR in several ways. It interrupts the fluency of the speech making the search system in the ASR difficult. Research to improve filled pause detection is mainly focused on speech feature extraction and modeling. A group of researchers have focused on acoustical feature extraction, including Mahesha & Vinod (2012); Kaushik et al.,(2010); Audhkhasi & Angeles (2009); Gaurav & Nigel, (2006), Goto et al., (1999). Another group of researchers have combined acoustical features with modeling techniques Zgank & Maucec, (2010); Li et al., (2008); Stouten et al.,(2006); Furui et al., (2005);. Although filled pause research is developing and improving, it still needs improvement as it has problems in discriminating between the filled pause and the elongated word (Veiga, 2011; Li, et al., 2010; Li et al., 2008).

2.3 PRE-PROCESSING

Speech pre-processing is a crucial step in any ASR development (Rosdi & Ainon, 2008). The basic pre-processing of speech signal consists of noise cancellation, pre-emphasis, framing, windowing, and voice activity detection (VAD) (Hariharan et al., 2012). Each step serves different purposes and is important to the whole process of speech recognition. For noise cancellation, there are several methods that can be implemented such as spectral subtractions and adaptive noise cancellation (Nilsson, 2002). The pre-emphasis is applied to compress the speech signal spectrally. High-pass filter is one of the techniques used (Mahesha & Vinod, 2012). The speech signal is sent to the high-pass filter as equation (2.1).

$$S_2(n) = s(n) - a * s(n-1) \quad (2.1)$$

where

a = 0.9 to 1

$S_2(n)$ = output signal

$s(n)$ = input signal

The z-transformed of the speech signal is $H(z) = 1 - az^{-1}$.

Framing process is executed to ensure the speech signal is chunked into smaller frames. Speech signal is non-stationary meaning that its statistical properties are not constant over time. However, speech is considered stationary if the frame size is within 20ms to 40ms (Singh et al., 2012). In most speech processing researches, the 20ms is chosen as the frame duration as can be seen in Verkhodanova & Shapranov, (2014); Reddy et al., (2013); Meduri et al., (2011); Zolnay & Haeb-Umbach (2006). The framing process is applied to create a series of frames with a wideband spectrum that is suited for capturing temporal changes in the speech signal. The short frames have narrow width and provide better temporal resolution (Ricke, 2006). It is also a norm that the speech is analyzed in short time frame rather than the whole speech signal at once. Windowing is important to reduce spectral artifact from the framing process (Proakis & Manolakis, 1996). Normally, windowing is applied after the framing process (Mahesha & Vinod, 2012; Hariharan et al., 2012; Singh et al., 2012; Marciniak & Krzykowska, 2012).

An additional pre-processing step can also be found in the literature especially in overcoming the speakers' volume variability. Volume or amplitude normalization is one of the techniques employed in ASR pre-processing. This technique has been found as early as 1986 in Jain, (1986). Various method of amplitude normalization were implemented such as radius of gyration method (Ye et al., 2002), mean and standard deviation method (Hariharan et al., 2012), and mean method, (Madsack, 2006). As a conclusion, the preprocessing of this research utilized framing, windowing, pre-emphasis by high-pass filter, threshold-based VAD method, and amplitude normalization.

2.3.1 Voice activity detection (VAD)

A review on ASR preprocessing was done by Singh et al., (2012). In their review paper, voice activity detection (VAD) is discussed as one of the important pre-processing stage in ASR. VAD is important if the processed speech is corrupted with background noise and to remove silence. Various techniques of VAD have been presented. The most widely used is VAD based on energy and zero crossing rates as thresholds. Basically, the principle of VAD is to extract features from a framed speech of (5 to 40)*ms*. A threshold limit is usually set from the noise only frames of the input signal. The extracted features are then compared with threshold to compute the VAD decision. An input frame is decided as voiced region if it exceeds the estimated threshold value. Otherwise, absence of speech is declared.

Meduri et al., (2011) used zero crossing count, speech energy and correlation between adjacent speech samples, first predictor coefficient from linear predictive coding analysis and the energy in the prediction error. In their research, these features are used in two categories of VAD methods which are threshold-based VAD that was formerly found by Rabiner & Sambur, (1975) and pattern recognition based-VAD. In this thesis, threshold-based VAD is chosen s it is a simple, well-established method (Atal & Rabiner, 1976) and is widely used until today. In Sakhnov et al., (2009), energy features such as linear energy and adaptive linear energy were exploited to gather the VAD threshold. Energy feature also was used in Aibinu, et al., (2011) as a threshold to evaluate the effect of VAD on isolated speech recognition of Yoruba

language. They concluded that VAD is important in their word recognition as the VAD significantly reduced the size of the speech sample prior to the recognition.

In this research, VAD is important to get the exact voiced segment of the filled pause and elongation. An inaccurate voiced segment extraction will result in lower classification performance. In Nemer et al., 2001, a voice detection algorithm based on the higher order statistics of speech is introduced. This algorithm utilizing the higher order properties of the speech that are distinct from those of Gaussian noise is reported to show significantly better performances than G.729B. However, when the background noise is neither Gaussian nor Gaussian-like, the higher order properties of speech and environmental noise is difficult to distinguish, thus leading to a performance degradation. Hence, the higher order properties alone is not adequate to separate it from noise. An improved VAD using higher order properties is proposed by Li et al., 2005 by combining the higher order properties with low band to full band energy ratio (LFER). It is proven in their research that the combination of higher order properties and the energy feature has improved the detection of unvoiced speech. Marciniak et al., (2012) has also used higher order properties for VAD. It was found that, the algorithm not only successfully detect the silence at the end and beginning of the word, but also silence in the middle of the word.

2.4 SPEECH FEATURE EXTRACTION

Feature extraction is one of the most important processes in automatic speech recognition. It is implemented to get the feature representation of the speech prior to the classification or recognition. The feature representation of the speech is very important to ensure the successful of the speech modelling (Li & Stern, 2003). There are various feature extraction techniques involved in speech modelling such as power spectral analysis (PSA), linear predictive coding (LPC), perceptual linear prediction (PLP), Mel frequency cepstral coefficient (MFCC), relative spectra filtering of log domain coefficients (RASTA) and first order derivative (delta). The most popular Mel frequency cepstral coefficients (MFCC) and linear prediction coding coefficient (LPCC) features were extensively used in the area of ASR.

MFCCs are based on the log spectral envelope of the speech signal, transformed to a non-linear frequency scale that roughly corresponds to that observed in the human auditory system. This representation is smoothed by applying a discrete cosine transform, resulting in a cepstral representation (Cucu, 2013). The MFCCs is a perceptually based Mel-spaced filter bank processing of the Fourier Transform. In Hu (2009), the used of MFCC as the feature vector in their filled pause and speech/ non-speech detection is one of the notable successful factor of their work. The popularity of MFCC as a feature vector also can be found in the work done by Rosdi & Ainon, (2008) for isolated Malay word recognition. An advanced research of Malay isolated word recognition which covers the spontaneous speech has been done by Seman et al., (2010). MFCC also used in Al-Hadad et al., (2008) for isolated Malay digit recognition. In their research, MFCC is chosen as the feature vector due to its sensitivity of the low order cepstral coefficients. Wu & Yan, (2004) also chose MFCC as a feature extraction in their filled pause work because it has been proven useful for speech recognition. One of the advantage of MFCC as found in their research is due to its stability towards lengthening. Lengthening factor is of importance in filled pause detection since it is one of the features of filled pause (Wu & Yan, 2004).

Another feature extraction technique that is similar to MFCC is perceptual linear prediction (PLP) (Rosdi & Ainon, 2008). An auditory-inspired cube-root compression is introduced in PLP technique. All-pole model is used to smooth the spectrum before the cepstral coefficients are computed. The PLP analysis is an extension of the Linear Prediction Coding (LPC) technique, but it is more effective because it takes advantage of some characteristics derived from the psycho-acoustic properties of the human ear (Peddinti & Hermansky, 2013). These characteristics are modelled by a filter-bank. LPC has been used in the early 1900 by Snell & Milinazzo, 1993. The outcome of linear prediction coding is a set of coefficients describing a filter, and a residual signal that is addressed as LPCC (Chetouani et al., 2009). In 2012, a comparison was made to classify speech disfluency and normal words occurring in stuttering speech by using MFCC and LPCC feature vector (Mahesha & Vinod, 2012). Based on their results, LPCC outperforms MFCC for both Linear Discriminant Analysis (LDA) and Artificial Neural Network (ANN) classifiers. For certain features such as formant frequency and fundamental frequency, the LPC method is a well-established method used. Formant frequency and fundamental

frequency is among the very well-established acoustical feature representing filled pause characteristic. Therefore, in this research, LPCC and MFCC were used to get different filled pause feature which are to get the formant frequency and MFCC.

One of the ways of executing filled pause researches is by using the acoustical-based rules. The filled pause researches that were done by rule-based acoustical algorithm based on different languages such as English, Mandarin, Dutch, Polish, Russian and Japanese are shown in Table 2.4.

Table 2.4:
Acoustical rule-based filled pause researches

Author/year/language	Features used	Performance measurement
Verkhodanova & Shapranov, 2014 (Russian)	Duration	Accuracy 80%
Karpinski, 2013 (Polish)	Fundamental frequency Formant frequency	Analysis on the feature only- no detection or classification
	Fundamental frequency Duration Shimmer	
Veiga, 2011 (Portuguese)	Fundamental frequency Energy 12-Mel frequency cepstral coefficients	Analysis on the feature only
Zuraidah & Knowles, 2006 (English and Malay)	Formant frequency (4-level) Fundamental frequency Energy Duration	Accuracy 80%
Ogata et al., 2009 (Japanese)	Duration Fundamental frequency Spectral envelope	Precision 92% Recall rate 85%
Audhkhasi & Angeles, 2009 (English)	Duration, Formant frequency (2-level) Fundamental frequency Mel Frequency Cepstral Coefficient (MFCC)	Precision 86%, Recall rate 45%
Gaurav & Nigel, 2006 (English)	Duration, Fundamental frequency Energy	Precision 86%, Recall rate 45%
Goto et al., 1999 (Japanese)	Duration Fundamental frequency	Precision 92% Recall rate 85%
Gabrea et al., 1999 (English)	Fundamental frequency Duration Energy Cepstral distance	Accuracy 96% Recall rate 93% False alarm 4%

Several well-established features were used by the filled pause researchers and the most popular acoustical feature are fundamental frequency (F0), speech energy, formant frequency (FF), duration and Mel-frequency cepstral coefficients (MFCC) and zero crossing rates (ZCR). Although filled pause is language dependent, many

cross-linguistic studies have shown that there are great similarities with regard to disfluencies across languages (Eukland, 2001). These well-established acoustical features were observed due to the minimal intonation produced while uttering the filled pause. When the filled pause is uttered with minimal intonation, the vocal tract characteristics remains unvaried (Audhkhasi & Angeles, 2009). Therefore, constant energy of the speech is produced that yield to stable formant frequency and flat fundamental frequency. Each of the acoustical features that were used in previous acoustical rule-based researches is discussed in the following subsections.

2.4.1 Fundamental frequency (F0)

Table 2.4 clearly shows that F0 is the most popular acoustical feature used among the researchers. Some of the researchers (Gabrea & O'Shaughnessy, 2000; Audhkhasi & Angeles, 2009; Kaushik et al., 2010) used the same acoustical features of F0. It was also found that flat frequency produced due to unvaried vocal tract characteristic during filled pause utterance is helpful in distinguishing filled pause among normal words. Goto et al., (1999) built another rule-based algorithm for filled pause classification. In their research, fundamental frequency (F0) along with spectral envelope is used to detect filled pause. Two thresholds were set by calculating the steadiness value of fundamental frequency and spectral envelope deformation of the filled pause utterances to build a filled pause detection algorithm. They successfully classified filled pause among normal words by using F0 with 84.9% recall and 91.5% precision rates. F0 is also utilized in Liu et al., (2006) and Ogata et al., (2009). The research is extended to observe the impact of filled pause classification prior to the ASR. It was proven in their research that the ASR performance has improved by ~3% accuracy when the filled pause classifier is included in their ASR system. The researchers have concluded that fundamental frequency of filled pause is flat compared to normal words. During filled pause and elongation production, there is very minimal articulation in the vocal tract since subsequent word has not yet been formed. Thus, this phenomenon produced a flat fundamental frequency. Verkhodanova & Shapranov, (2014) has implemented rule-based method in the filled pause and elongation detection of Russian language. In their research, F0, duration and formant frequency are used. A threshold is set on each acoustical feature. They

successfully achieved 80% accuracy for filled pause and elongation. In their research, the elongation occurs in two types of voiced and unvoiced fricatives. They concluded that the aforementioned extracted acoustical features is not suitable to represent unvoiced fricative elongation.

2.4.2 Energy

Energy of speech is another acoustical feature utilized in filled pause research. The unvaried vocal cord during filled pause pronunciation produced constant energy (Gaurav & Nigel, 2006; Li et al., 2008). This acoustical feature is extracted in a short time basis of the speech and called as Short Time Energy (STE). STE is shown to be able to classify voiced and unvoiced speech segment effectively (Jalil, Butt, & Malik, 2013). This feature is also used in Gaurav & Nigel, (2006); Kaushik et al., (2010); Li et al., (2008); Medeiros, Moniz, et al., (2013); Veiga, (2011); Zgank & Maucec, (2010) for filled pause detection. It was found that energy of the filled pause is more stable and constant along the filled pause compared to normal words due to the unvaried phoneme pronunciation (Gaurav & Nigel, 2006).

2.4.3 Formant Frequency (FF)

Formant frequency is produced in relation with vocal tract activity. The oral opening from the larynx to the lips and the joined nasal and oral passages produces the vocal tract. Numerous different patterns can be produced by the vocal tract depending upon the shape and movement of the tongue, mouth, teeth, lips and jaw. Based on this physiology, the vocal tract produces frequency shaping from the larynx. Eventually, a new source for sound production, which is also known as impulsive source, is generated.

The vocal tract can be considered like a linear filter with resonances due to the relation between glottal airflow velocity input and vocal tract airflow velocity output in certain situations. These resonances produced from the vocal tract are called formant frequencies (Thomas, 2008). In a spectrogram view of speech waveforms, the peaks of the vocal track response that are viewed as time invariant with all-pole linear system are parallel to its formant frequencies (Kammoun et al., 2006). The value of

formant frequency is proportional to the shape of the vocal tract during utterance production. For example, a vowel utterance has a different position of jaw, teeth, lips and tongue in comparison with a consonant, thus producing different formant frequencies.

Formant frequency is also one of the common acoustical features used in the filled pause researches as early as 2004 (Wu & Yan). The level of formant frequencies used is important as each level represent different location of speech articulation (Delattre et al., 1952). Audhkhasi & Angeles, (2009) used two level of formant frequencies (i.e. FF1 and FF2) to build a detection algorithm for filled pause detection among normal words. A threshold that is calculated from the stability measure of the formant frequencies are used in the rule-based algorithm development. They found that these two formant frequencies performed better when used to classify filled pause compared to Mel-frequency cepstral coefficient and fundamental frequency. On the other hand, Kaushik et al., (2010) suggested that the first level of formant frequency was found as not sufficient to represent filled pause. Based on their observation, some filled pauses and normal words showed the same pattern of first formant frequency, thus causing filled pause misclassification. Therefore, four levels of formant frequency are implemented and they found that the second, third and fourth level of the formant frequencies are better representations for the filled pause. The fourth level of formant frequency along with fundamental frequency and speech energy are used to develop a rule-based filled pause classification algorithm. Their algorithm successfully achieved 80% accuracy in filled pause classification.

2.4.4 Duration

Duration of filled pause is the other acoustical feature that has almost similar range among languages. The repeated vowels in the filled pause is longer in duration compared to vowels in normal words (Kaushik et al., 2010). In Li et al., (2008) the duration that is defined for Mandarin language filled pause is above 200 milliseconds (*ms*). For Portuguese language, a duration threshold of 350*ms* is set to analyse the filled pause characteristics. Merely the same threshold for filled pause of 355*ms* is also used by Bartkova, 2005 to analyse French language filled pause. Veiga, (2011) performed an acoustical analysis on filled pause and elongation of Portuguese

language. F0, energy, duration and FF were extracted. Based on their observation, these extracted acoustical features did not show significant differences between filled pause and elongation.

2.4.5 Mel Frequency Cepstral Coefficients (MFCC)

Mel frequency cepstral coefficient (MFCC) is a feature commonly used in many speech processing researches including filled pause work. MFCC represents the speech into its equally spaced frequency based on Mel-scale which is close to the human auditory system response (Holmes, 2001). In ASR, MFCC is a common representation of the speech signal (Shaneh & Taheri, 2009). MFCC has been widely used in filled pause research (Mahesha & Vinod, 2012; Veiga, 2011; Zgank & Maucec, 2010; Audhkhasi & Angeles, 2009; Stouten, 2008; Stouten et al., 2006; Wu & Yan, 2004; Stouten & Martens, 2003).

2.4.6 Zero Crossing Rates (ZCR)

Zero crossing rates (ZCR) is a well-known speech acoustical feature. Apart from applying Fourier transform on the speech signal, ZCR provide simpler and less computational cost in order to provide the spectral information of the speech (Deng & O'Shaughnessy, 2003). ZCR is extensively used in ASR researches since 1971 (Ito & Donaldson, 1971) mostly for recognition of voiced and unvoiced speech segment purpose (Bachu et al., 2008; Radmard et al., 2011; Jalil et al., 2013; Taher et al., 2014;). ZCR is implemented in non-speech analysis as seen in Cai et al., (2003) to detect speech highlight such as laughter, applause and cheer. However, ZCR is hardly to be utilized in the specific filled pause detection research except in Li et al., (2010).

In this research, investigations of well-established acoustical features are conducted to identify the most suitable feature(s) for Malay language disfluency. All the six acoustical features commonly used in disfluency-related researches found in the literature are further tested and explained in this thesis. They are F0, FF, energy, duration, MFCC and ZCR.

2.5 PATTERN CLASSIFICATION

Classification can be divided into supervised and unsupervised classification. The implementation of supervised classification in this research is due to the need to train the labeled data before testing. Supervised classification is a process to map an input to its own category based on the labeled training data. Various supervised classification algorithms have been proposed in the literature such as artificial neural networks (ANN), support vector machines (SVM) and Bayesian classifiers.

In automatic speech recognition, methods such as ANN (Nie & Zeng, 2004), (Seman, and, & Bakar, 2010), Hidden Markov Model (HMM) (Rosdi & Ainon, 2008; Al-alaoui et al., 2008), linear discriminant analysis (LDA) (Hariharan et al., 2012), SVM (Khan, et al., 2011; McLaren et al., 2009), Gaussian mixture model (GMM) (Wu & Yan, 2004; Marciniak & Krzykowska, 2012) and naïve Bayes (Sanchis et al., 2012; Tóth et al., 2005; Ye et al., 2002) were used. Among the aforementioned classifiers, HMM that was pioneered by Rabiner & Juang, (1986) has an impressive history in ASR for the purposes of classification, segmentation, and clustering.

A HMM is comprised of a set of states Q , a set of transition probabilities A , a set of observation likelihoods B , a defined start state and end state(s), and a set of observation symbols O , which is not drawn from the same alphabet as the state set Q (Jurafsky & Martin, 2007). The advantage of HMM-based modeling is it only requires a small amount of training data. HMM has shown a high reliability of performing speech recognition. The development of ASR improved progressively since the innovation of HMM in 1986's for English language digit recognition. HMM in speech recognition studies has been applied for many purpose to improve the accuracy of ASR.

Although HMM is popularly used in ASR, many improvements of HMM are made by researchers. In Yuan, (2008) an attempt to solve the defects of duration modeling of classical HMM was made. The Markov Family Model (MFM) was introduced by replacing independence assumption with conditional independence assumption with the fact that the assumption of successive observations is independent and identically distributed within a state is unrealistic. As a result, the word error rate (WER) was reduced by 1.4 % in comparison to the traditional HMM. In Han (2003), a comparison was made by developing HMM-based speech

recognition Integrated Circuit (IC) with baseline HMM ASR software. The invention has the ability to reduce the circuit complexity with the same accuracy of the HMM ASR software. A system of HMM-based English digit recognition was developed by Gunawan et al., (2011), and achieved promising recognition rate of 99.5%. The researcher discovered that HMM performed better when applied in clean environment. However, the data that is used in this research contains high noise as it was recorded live in a debate session of Malaysian Parliament. Therefore HMM classification is not suitable in this research.

Researchers have also introduced hybridization between HMM and other classifiers such as SVM, GMM, and ANN to improve the HMM performance. A Gaussian Mixture Model (GMM) is a parametric probability density function denoted as a weighted sum of Gaussian component densities. GMM is of assistance for classification tasks that don't require the importance of temporal information such as speech and music classification or speech and non-speech detection (Hu, 2009). However, as stated in Kruger et al., (2006) GMM cannot discriminate the speech unit very well as it is trained by likelihood maximization. Kruger et al., (2006) has proposed the integration of SVM parallel mixture pioneered by Collobert, (1920). Likelihood maximization method assumes the correctness of the models and thus suffers from poor discrimination (Bouclard & Morgan, 1998). To overcome this limitation, they used SVM as a model of emission probabilities and hybrid with HMM. Results showed that the hybrid classifier between SVM and HMM are better in performance compared to GMM and HMM. However, the hybridization between SVM and HMM is highly time consuming caused high computation time as found in Yang, Wang, & Sun, (2010). Instead of relying on the hybridization method, the weighted autoregressive HMM (WARHMM) was exploited in the research to solve the problem of features vectors independent assumption and small usage of datasets.

In Revathi (2012), the vector quantization (VQ)-based speech recognition is compared with VQHMM-based speech recognition. Result indicated that the hybridization of VQ and HMM outperformed the standalone VQ-based ASR with difference of 7% accuracy for isolated digit and 1% for continuous speech. It can be seen that, for a better performance of HMM, more complex modification need to be taken.

Another filled pause researches group used the combination of acoustical rule-based and classification. The filled pause work combined rule-based and classifications as summarized in Table 2.5. According to Li et al., (2008) acoustical rules alone are not effective for filled pause classification when different data is used since rule-based approach is domain-specific. The combination of acoustical rules algorithm and classification was done in Moniz et al., (2009). In the research, F0 and Classification and Regression Tree (CART) were used to classify European Portuguese language filled pause. At first, CART was used to separate the speech into normal words and disfluencies (filled pause and elongation). The pattern observation on the disfluencies was then evaluated by analyzing the pattern of the fundamental frequencies, and achieved 78.3% classification accuracy. Although F0 is popularly known as flat, Moniz et al., (2009) found that some of their filled pause and elongation presented ascending patterns.

Another classifier that was used in filled pause classification is Hidden Markov Model (HMM). In Li et al., (2008), acoustical features of F0, STE and MFCC are utilized to detect Mandarin language filled pause. The F0 and STE are used as a filtering mechanism before the classification process to eliminate non-filled pause segments of the speech. The earlier sign of non-robustness of F0 and STE in classifying filled pause and elongation can be noticed as only normal words are filtered. Large portion of filled pause and elongations are observed after the filtering process. For classification, MFCC along with STE and F0 are used to classify the filled pause by using HMM. They successfully achieved 80.66% and 92.59% precision and recall rates for the filled pause class. However, there is no specific precision and recall rate for the elongation and normal words. They concluded that the misclassified filled pause is due to the occurrences of elongation.

In Stouten, (2008), MFCC, spectral envelope stability and duration are extracted and evaluated to detect filled pause using Multilayer Perceptron (MLP) classifier. A Gaussian Mixture Model (GMM) is applied prior to the MLP classification to eliminate the non-filled pause segments. They successfully achieved 74.1% recall and 83.7% precision rates in the filled pause detection. GMM was also used in Wu & Yan, (2004). Wu & Yan employed MFCC, linear predictive coding coefficient (LPCC), Formant Frequency 1, 2 and 3 to classify filled pauses by using GMM. They successfully achieved 86.6% detection rates.

Table 2.5:
Filled pause research based on acoustical and classification method

Author/year	Feature used	Classifier	Performance measurement
Medeiros et al., 2013 (European Portuguese)	Fundamental frequency Energy Duration	CART Naïve Bayes Logistic regression Multilayer perceptron J48	Recall 99% Precision 97% F-Measure 98%
Mahesha & Vinod, 2012 (English)	Mel Frequency Cepstral Coefficient	k-NN	Accuracy 87%
Zgank & Maucec, 2010 (Slovenian)	Energy MFCC	HMM	ASR Accuracy 65%
Moniz et al., 2009 (European Portuguese)	Fundamental frequency	CART	Accuracy 78%
(Li et al., 2008)	MFCC Fundamental frequency Energy	HMM	Precision 80.66% Recall 92.59%
Stouten et al., 2006 (Dutch)	Duration Fundamental frequency MFCC	MLP-ANN	Precision 84% Recall 51%
Nunes & Neves, 2006 (Portuguese)	-	Language Model	False alarm 50%
Wu & Yan, 2004 (Mandarin)	Formant frequency MFCC	GMM	Accuracy 87%
Peters, 2003 (English)	-	Language Model	ASR error rate reduced to 2%
Stouten & Martens, 2003 (Dutch)	MFCC Duration	MLP-ANN	Recall 74% Precision 84%
Shriberg et al., 1997 (English)	Duration Fundamental frequency	Language Model CART	Accuracy 90% Recall 92% False alarm 12.9%

Different classifier such as classification and regression tree (CART), naïve Bayes, logistic regression, Multilayer perceptron and J48 were used to classify normal word and disfluencies (Medeiros, Batista, et al., 2013). The disfluencies are inclusive of filled pause, repetition, restart and elongation. They have successfully achieved a recall rate of 99.7 %, precision of 96.80% and F-measure of 98.3% by using CART. The CART performance also is tested in (Shriberg et al., 1997) with the same

acoustical features used in Medeiros et al., (2013). They achieved accuracy of 89.70%, recall of 92.3% and false alarm of 12.9%.

In Stouten & Martens, (2006), MLP of Artificial Neural Network (ANN) was used. The aim of using MLP-ANN in their research is to predict the posterior probability of a segment is having filled pause type of speech. However, the MLP-ANN training produced poor performance of classification since the prior probability is unknown. A pre-processing step needs to be taken in their research to isolate the non-FP segments from their database by combining GMM with the MLP. Mahesha & Vinod, (2012), used k-NN to classify speech segments into normal and disfluencies by using MFCC as acoustical feature and achieved 86.67% accuracy.

Popular classifiers such HMM (Li et al., 2008), k-NN (Mahesha & Vinod, 2012), CART, naïve Bayes, logistic regression, MLP-ANN and J48 Medeiros et al., (2013) were formerly used in filled pause classification. Although some of the research used rule-based approach only while others implemented classification method, both researches has its own importance. Although it was claimed by Li et al., 2008 that the rule-based approach is not sufficient for filled pause classification, the features that made up the rules are importantly needed to be extracted prior to the classification stage. It was proven in the previous research that an irrelevant features lead to misclassification of filled pause. Therefore, a relevant acoustical features need to be carefully extracted and tested for better filled pause classification.

The previous filled pause researches of Table 2.4 and Table 2.5 are broadly classified into two categories which are classification of filled pause and elongation as one class and classification of both filled pause and elongation into two separate classes. The occurrence of elongation is one of the factors which degrade the performance of filled pause classification as they shared the same acoustical features. To overcome this limitation, most of the researchers grouped them into one class on the basis that they serve the same linguistic meaning. However, the classification of filled pause and elongation into separate classes are mutually agreed (Li et al., 2008, 2010; Veiga, 2011). Elongation has its own linguistic meaning. Therefore, detecting it as filled pause disrupts the semantic of the speech sentence. Thus, this research focuses on the filled pause and elongation classification as two separate classes.

Wu & Yan (2004) modeled the two levels of FF using Gaussian Mixture Model (GMM). The filled pauses are then classified using Karhunen Loeve Transform (KLT) that is also known as Principal Component Analysis (PCA) (Hanilçi & Ertaş, 2009) and LDA. Results showed promising classification percentage of 86.6% using LDA classifier. In the occurrence of several features, however, observation probability functions of HMMs denoted conditional dependence among features given the state (Avilés-Arriaga et al., 2011). This makes it difficult to visualize independence relationships of features and their statistical behavior. Instead of using HMM, the utilization of naïve Bayes lessens the dependence assumption between features (Avilés-Arriaga et al., 2011).

Naïve Bayes is one of the classifiers that have received a lot of attention in ASR (Sunny & Jacob, 2013; Sanchis et al., 2012; Tóth et al., 2005). Naïve Bayes has been used by Medeiros et al., (2013) in speech disfluency research. It is a simple and efficient probabilistic classifier that assumes independent relation between features in the classification. However, in reality this assumption is almost unachievable (Panda et al., 2010). Although it ignores cooperative effects of input features (unless cross-features are computed before the implementation), Naïve Bayes is known as an efficient machine learning method that works well for different classification tasks (Mitchell, 1997). In the filled pause and elongation classification circumstance, the dependency degree among features is not fully understood. It is well-known that filled pause and elongation have the same acoustical features such as energy, fundamental frequency and formant frequency (Stouten et al., 2006; Li et al., 2008; Veiga, 2011). Hence, the choice of classifier that has the ability to classify the filled pause and elongation regardless the features dependability is an extra-advantage and worth attempting. Therefore, Naïve Bayes is selected as the classification for this research.

2.5.1 Naïve Bayes

Naïve Bayes theorem is important in terms of its computation simplicity and provides direct multi-group feature classification. Naïve Bayes classifier was developed in 1702 by Thomas Bayes (Altwaijry, 2013). At present, it is still being used and its performance is proven in many fields of researches (Altwaijry, 2013) including speech recognition (Iqbal, 2014; Tóth et al., 2005). Naïve Bayes classifier is

also a supervised learning algorithm (Iqbal, 2014). In supervised learning the aim is to train a classifier to map the input to output given the correct values provided by the controller (McHugh, 2000). Despite its simplicity, Naïve Bayes classifier has successfully addressed many research area problems including speech recognition, medical and biological (Sheppard, 2013) especially when many features are to be used for modeling. The Naive Bayes classifier remains to be a widely used learning algorithm for data mining applications due to its simplicity and linear run-time (Hall, 2007). Naïve Bayes is also a fast-supervised classification technique which is suitable for large-scale prediction and classification tasks on complex and small data. Zaidi et al., (2013) stated that Naïve Bayes classification is a simple probabilistic machine learning algorithm, making it less complex and more efficient. In Wang & Zhang, (2005) an automatic text classification was built using Naïve Bayes. They have successfully achieved the highest precision of 80% and a recall rate of 74% in the text classification. In Sanchis et al., (2012) Naïve Bayes classifier was used to classify word into its respective class. Several types of features inclusive of acoustical, language model, decoding process features were utilized. The lowest error rate of 6.6% was achieved in their research.

It is well-understood that Naïve Bayes is a well-known classifier. However Naïve Bayes is also associated with fusion classification whereby the information of different types of features is combined using posterior probability product calculation to decide the final class of the input. Naïve Bayes fusion is also called as statistical fusion (Cremer et al., 2001). It was reported that, fusion technique for classification introduced in ASR research enhanced the performance compared to classification of single feature (Planet & Iriondo, 2012). Fusion methods can be viewed in different ways; however, there is no clear and strict classification of fusion methods. Previously, the outputs of each single feature classifier were fused using several methods such as fuzzy integral, sum, product, max, min, weighted arithmetical mean and Naive Bayes theorem (Temko et al., 2008). Naïve Bayes is also addressed as scores fusion classifier (Trabelsi & Ayed, 2013), decision fusion or classifier fusion (Cremer et al., 2012) and product rule (Altınçay, 2005). In the sense of simplicity, naïve Bayes fusion is the best as it requires small amount of information that need to be combined (Trabelsi & Ayed, 2013).

It was reported that, the advantage of each feature can improve the classification by combination of techniques (Castanedo, 2013). Several combination techniques were implemented in the previous researches such as Naïve Bayes, Fuzzy Logic, ruled-based fusion, and Dempster-Shafer Theory (Cremer et al., 2001). Naïve Bayes allows the combination of different features and has become one of the famous techniques in speech research (Makkook, 2007; Lee & Park, 2008 and Sunny et al., 2013). It is also proven as computationally more efficient than the other techniques (Cremer et al., 2001). Naïve Bayes assumes the features are conditionally independent given the class labels (Xu et al., 1992). However, according to Domingos & Pazzani (1997), the Naïve Bayes classifier retains its good performance even though the independence assumption is broken. According to Zaidi et al., (2013), features are known to be independent if they are used individually as training features by different classifier.

Generally, estimating the probability density is part of the processes in naïve Bayes classification. Estimating the conditional probability density is important for Bayes classifier to study the pattern of the data (Sunny et al., 2013). The estimation of the probability density will provide knowledge of how the data is distributed according to its own class. The probability density estimation is then used to obtain the conditional probability of each class of data prior to the posterior probability calculation.

Conditional probability $P_i(X_i | C = c)$ is the probability of feature value in the i^{th} position is equal to x_i with a given class; c . Density estimation involves mathematical and statistical calculation (Donthu, 1991). Various techniques are available for density estimation in naïve Bayes classification. Probability density function estimation approaches are divided into parametric, non-parametric and semi-parametric. The most popular approach by Bayesian researcher is estimating the true density of the variables using a parametric approach, namely Gaussian function. Although Gaussian function may provide a reasonable approximation to many real-world distribution, it is usually not the best approximation (Pérez et al., 2009). The Gaussian function assumes the continuous variable or data as per Gaussian distribution while many real data may not be represented in the Gaussian distribution (Soria, et al., 2011). Gaussian density function is estimated by obtaining the mean and variance of the data as shown in equation (2.2).

$$P_i(x_i | C = c) = \frac{1}{N_c h} \sum_{j=1}^{N_c} K(x_i, -x_{j|i} | c) \quad (2.2)$$

where,

$x_{j|i}$ = is the feature value in the i -th position of the j -th input $X=(x_1 x_2 \dots x_i \dots x_n)$ in class c .

h = is a bandwidth, or a smoothing parameter.

N_c = is the number of data X belonging to class c in the training set.

$K_H(\cdot)$ = is the kernel function used.

The probability density function can also be estimated by using non-parametric approaches which is called Kernel Density Estimation (KDE) (Silverman, 1986). KDE and Naïve Bayes classification technique has been used in many classification areas. In Murakami & Mizuguchi, (2010) Naïve Bayes classification was implemented with KDE on the training data to detect the protein-protein interaction sites. The implementation of KDE in the Naïve Bayes classification has shown improved classification performance as compared to smoothing technique for probability density estimation. The KDE-based probability density estimation Naïve Bayes classifier has achieved 23.3% higher accuracy as compared to without KDE. KDE is also beneficial for a highly imbalance training data as the results show improved performance as compared to smoothing technique. Solà-soler et al., (2011) compared KDE and Gaussian estimation in their Naïve Bayes classification for sleeping syndrome classification. The study found that KDE performed 27% higher as compared to Gaussian in every case. Therefore, KDE is chosen as the probability density estimation in this research.

KDE has more flexible properties than Gaussian approach in order to fit non-Gaussian densities. Kernel density estimation is known as non-parametric distribution function estimation approach due to its characteristics in estimating the distribution without assuming any underlying distribution for the variable in the data set. A Kernel is a function that follows the subsequent properties (Silverman, 1986):

- i. non-negative
- ii. real-valued
- iii. even

iv. its definite integral over its support set must equal to 1

There are varieties of Kernel function which are inclusive of uniform, triangle, Epanechnikov, Quartic, Triweight, Gaussian and Cosinus (Altwaijry, 2013). The widely used and performing Kernel function, Gaussian Kernel with zero mean and variance (Pérez et al., 2009) is chosen in this research. The Gaussian Kernel function $K_H(\cdot)$ is estimated as in equation (2.3).

$$K(a, b) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - x_i)^2}{2h^2}\right) \quad (2.3)$$

where,

h = the Kernel bandwidth

x = the feature value

Although the Kernel Gaussian function is similar to the parametric-based density estimation that is Gaussian density estimation, both the functions are executed differently. In Gaussian density estimation the estimation is conducted based on statistical parameters such as mean, variance and standard deviation and by assuming the data are normally distributed. Meanwhile, the Kernel Gaussian function of non-parametric density estimation method, the probability is estimated discretely. Gaussian function is the most popular implementation in KDE as noted in Pérez et al., (2009); Murakami & Mizuguchi, (2010); Tanaka et al., (2014). The KDE function is defined as in equation (2.4).

$$P_i(x_i | C = c) = \frac{1}{N_c h} \sum_{j=1}^{N_c} K(x_i, -x_{j|i|c}) \quad (2.4)$$

where,

$x_{j|i|c}$ = is the feature value in the i -th position of the j -th input $X=(x_1 x_2 \dots x_i \dots x_n)$ in class c

h = is a bandwidth, or a smoothing parameter.

N_c = is the number of data X belonging to class c in the training set

$K_H(\cdot)$ = is the kernel function used

2.5.1.1 Bandwidth parameter selection for Kernel density estimation

The choice of Kernel function is not as important as bandwidth selection (Pérez et al., 2009; Guidoum, 2015). The selection of available Kernel function does not affect the shape of the Kernel estimator. Even by using different available function good results can be attained (Guidoum, 2015). Kernel estimator is considered as a sum of "bumps" placed at all data values. Figure 2.4 shows the bumps of the data and the estimated kernel density.

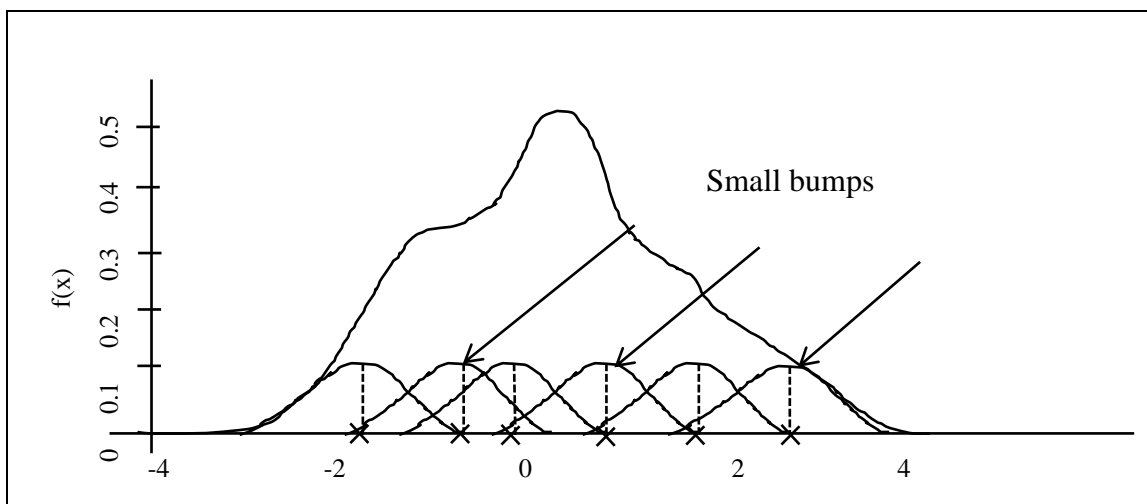


Figure 2.4: Kernel density estimation example (Silverman, 1986)

The kernel function, K determines the shape of the "bump" (Silverman, 1986) and the smoothing parameter h determines the width of the "bump". According to Zambom & Dias, (2012), small bumps produced by KDE represent noise which does not reflect the true estimated density. The value of the bandwidth functions as a scaling factor to provide the spread of the Kernel at each point. Several techniques are available for bandwidth selection in KDE-based Naïve Bayes. The most widely used is the Silverman's rule of thumb (Cowell et al., 2013) which is defined as equation (2.5).

$$\hat{h}_{opt} = 0.9 \min(\hat{\sigma}; \frac{\hat{q}_3 - \hat{q}_1}{1.349}) n^{-\frac{1}{5}} \quad (2.5)$$

where;

\hat{h}_{opt} = KDE bandwidth function by Silverman's rule of thumb.

$\hat{\sigma}$ = standard deviation of the data.

\hat{q}_3 = third quartiles calculated from the data.

\hat{q}_1 = first quartiles calculated from the data.

The other techniques involved in KDE bandwidth selection are common variation of factor 1.06 (Scott, 1992), biased and unbiased cross-validation (Stone 1974; Rudemo 1982; Bowman 1984), and pilot estimation of derivative (Sheather & Jones 1991; Ruppert et al., 1995). These techniques are not as simple as the technique defined by the Silverman's rule of thumb. These techniques require complex numerical calculation (Cowell et al., 2013). The subsequent process of the Kernel density estimation (KDE) is the selection of bandwidth smoothing parameter, H . H is a free parameter that controls the smoothness of the estimates. The selection of bandwidth parameter is crucial in Kernel density estimation. A very small H parameter causes the obvious spurious fine structure, while if H parameter is too large then the bimodal nature of the distribution is obscured (Murakami & Mizuguchi, 2010). Generally, the bandwidth parameter can be selected by using manual or automatic approach. In the manual approach, the bandwidth parameter can be done by randomly choosing the smallest to the biggest number until the optimal bandwidth is found. However, according to Silverman, (1986), this manual method is worrisome and incomplete. Instead, an automatic H parameter selection is more efficient (Hansen, 2009). Various techniques are involved in automated bandwidth selection, such as:

- i. Common variation of factor 1.06 (nrd) (Scott, 1992).
- ii. Silverman's rule of thumb (nrd0) (Silverman, 1986).
- iii. Biased (bcv) and unbiased cross-validation (ucv) (Stone, 1974; Rudemo, 1982; Bowman, 1984).
- iv. Pilot estimation of derivative (SJ-dpi) (Sheather & Jones 1991, Ruppert et al., 1995).

Each of these methods is tested on the acoustical feature and an example of bandwidth selection using ZCR is plotted in Figure 2.5. Based on the KDE of Figure 2.5, 'nrd', 'ucv', 'SJ-ste' and 'SJ-dpi' methods produces several small bumps due to noise. According to Zambom & Dias, (2012), small bumps produced by KDE represent

noise which does not reflect the true estimated density. On the other hand, bandwidth methods ‘nrd0’ and ‘bcv’ do not produce small bumps. Therefore, a more accurate KDE is derived. The ‘nrd0’ which is known as Silverman’s rule of thumb (Zambom & Dias, 2012) is the most commonly used (Murakami & Mizuguchi, 2010) bandwidth selection for KDE. Thus, it is chosen to automatically select the KDE bandwidth.

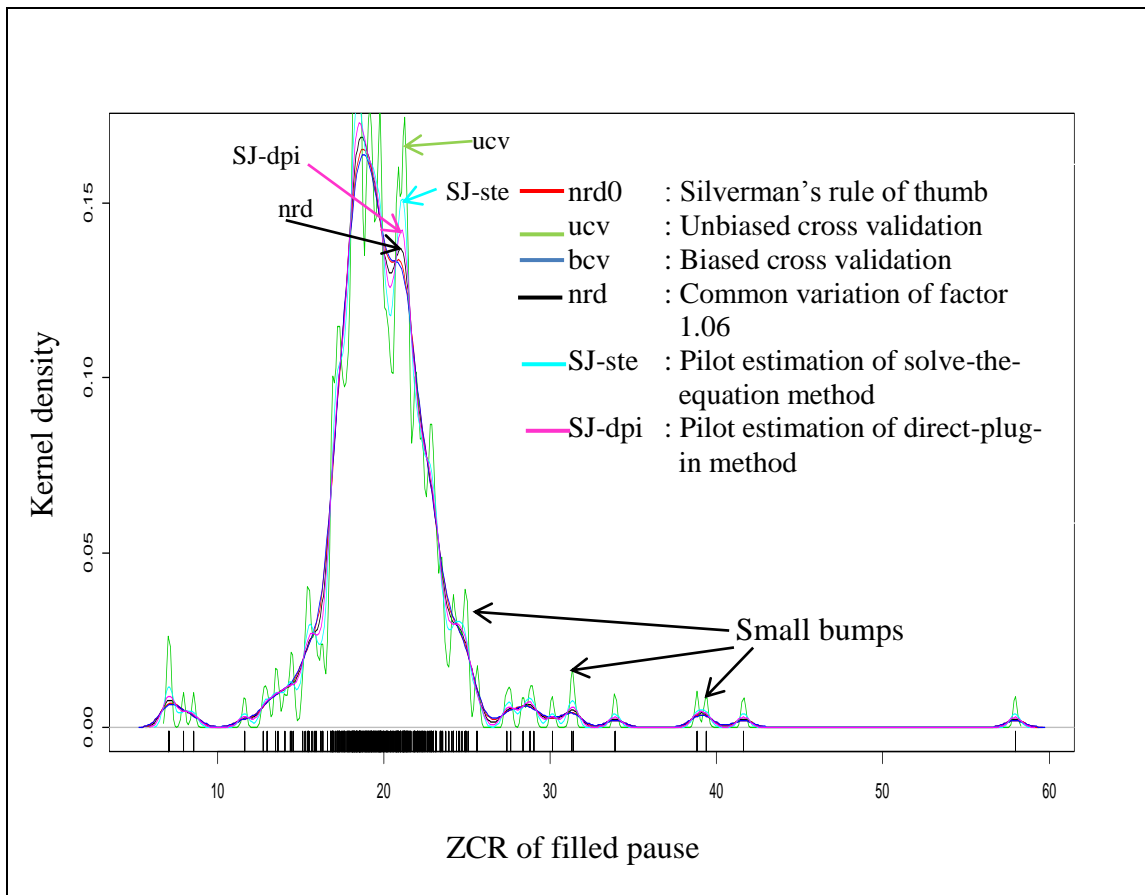


Figure 2.5: KDE plot of various bandwidth methods

2.6 EVALUATION OF CLASSIFICATION

Generally, the evaluation of a classifier can be measured by defining a matrix with the numbers of correctly and incorrectly classified class, known as confusion matrix. In this research, disfluencies are classified into two classes that are filled pause and elongation. In a two class-classification, the confusion matrix is defined as Table 2.6 (Costa et al., 2007).

Table 2.6:
Confusion matrix of classifier evaluation

	Predicted Class	
True Class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

where,

False positives (FP) : Test sample from negative class misclassified as positive class.

False negatives (FN): Test sample from positive class misclassified as negative class.

True positives (TP) : Test sample from positive class classified correctly.

True negatives (TN) : Test sample from negative class classified correctly.

Accuracy is the most common measurement chosen for classifier evaluation (Costa et al., 2007). It evaluates the effectiveness of the classifier by its percentage of correct classification. In disfluency classification, (Mahesha & Vinod, 2012) used accuracy to measure their classification of filled pause and normal speech. Accuracy is also used in Kaushik et al., (2010); Zgank & Maucec, (2010); Medeiros et al., (2013). Hu, (2009) used true positive and false positive to calculate the score of their classification performance. They stated that accuracy measurement is meaningless since there are big ratio of number between speech and non-speech (includes filled pause). However, in this research the number of filled pause and elongations are the same. By using accuracy, the number of correctly classified filled pause or elongation from the total amount of the disfluencies can be gathered. Therefore, the accuracy measurement is one of the evaluation methods chosen in this research.

There are several other existing evaluation methods exist in filled pause researches. The computation of the recall, precision, and F-measure are the most common evaluation methods in filled pause classification researches (Li et al., 2008; Audhkhasi & Angeles, 2009; Medeiros et al., 2013). Precision is a measure which estimates the probability that a positive classification is correct. In the view of filled pause classification, precision is defined as the ratio of total number of correctly classified filled pauses to the number of classified filled pauses. Therefore, it gives the insight of how many classified filled pause is correctly classified. *TP* and *FP* from Table 2.8 is used for precision computation given in equation (2.6).

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \quad (2.6)$$

On the other hand, recall rates is calculated as the ratio of total number of correctly classified filled pauses to the total number of filled pauses. Recall is used to evaluate the effectiveness of a classifier for each class in the binary problem (Costa, Lorena, C.P.L.F.Carvalho, & A.Freitas, 2007). The recall, also known as sensitivity or true positive rate, is the proportion of test samples belonging to the positive class which were correctly classified as positive. From Table 2.6, the TP and FN is used to calculate the recall rate as given in equation (2.7).

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|} \quad (2.7)$$

The value of precision and recall is then harmonically computed to get the F-measure. F-measure is a measurement that determines the precision and recall capabilities of a classifier that is used as a general guide to determine the overall quality performance of classifier. Therefore, the common evaluation method of recall, precision, F-measure, and accuracy are also used in this research.

2.7 RELATED MALAY LANGUAGE SPONTANEOUS SPEECH RESEARCHES

Most Malay language researches in Malaysia covered read speech as summarized by (Fook et al., 2012). Only a few researchers focused in spontaneous speech (Izzad et al., 2013; Chong et al., 2012; Seman et al., 2010; Zuraidah & Knowles, 2006; Pillai, 2006). Izzad et al., (2013) focused on recognizing voiced and unvoiced segments of the spontaneous speech. Although Seman et al., (2010) focused on spontaneous speech, the research is limited to endpoint detection of isolated words for ASR.

Very few research on Malay language disfluencies existed in the literature except in Chong et al., (2012); Zuraidah & Knowles, (2006). Pillai, (2006) stated that the spontaneous speech contain a normal phenomenon which is called as self-repair.

Self-repair includes inability to retrieve lexical items, and the incorrect use of pronunciation, lexis or syntax. When speakers cut their speech off in the midst of a word speakers may produce hesitation in their speech, such as filled pauses (e.g. *ah*, *ahm*, *er*), silent pauses and elongated segments (Pillai, 2006). Time intervals have been used in the research to detect error, interruption point, onset of the repair and editing phase. By investigating the time interval of each component in spontaneous speech, they have found about 264 of the utterances are interrupted by hesitation which included silent and filled pause as well as elongation.

Zuraidah & Knowles, (2006) investigated the types of prosodic features such as duration, pitch and tempo that involved in signaling the status of the speaker's whether to stop or to maintain their turn in a conversation. In their research, filled pause is considered as a verbal device that is used to prevent interruption from others while the current speakers plan his utterances. Although filled pause is one of the disfluencies studied, their research focused more on linguistic aspect of the filled pause and its function in spontaneous speech.

2.8 MALAY LANGUAGE SPEECH SOUNDS AND RULES

Malay language is an Austronesian language spoken by Malay people who are native to the Malay Peninsula, southern Thailand, Singapore, Brunei, and parts of Sumatra. In Malaysia, it is called Bahasa Melayu and it is the official language of Malaysia. Malay language is agglutinative in nature, meaning the words can be altered by adding the necessary affixes (Lee et al., 2013). Table 2.7 shows the examples of affixes affecting Malay words. From the table, it shows that the primary word 'susu' (i.e. milk), is transformed into verbs and new lexicon. The smallest unit that forms the affixes-added word or primary word in Malay language is a phoneme. The substitution of this unit with another might make a difference in meaning (Ting et al., 2001) such as the word "*buku*", if the second "*u*" is substituted with "*a*" the word will become "*buka*" (i.e. to open) or substitute "*k*" with "*l*" the word will become "*bulu*" (i.e. fur).

Table 2.7:
Examples of Malay word alteration and the transformation

Primary words	Prefixes	Suffixes	Infixes	Confixes
Susu (milk)	<i>Meny-usu</i> (is nursing) (verb)	Susu-an (milk) (nouns)	<i>Meny-usu-kan-</i> <i>Nurse</i> (verb)	<i>Ten-usu</i> (milk product) (new lexicon)

Generally, there are three major categories of Malay phonemes namely Vowels (V), Consonant(C) and other miscellaneous totaling to 36 Malay phoneme (Maris, 1966). The structure is comparatively similar with English language as presented in Figure 2.6 (Karim 1996).

The first category of phoneme, known as vowel comprises six vowels which are /a/, /e/, /i/, /o/, /u/. The vowel sound is produced during the expiration activity from the lunge and mouth without any noise and pronounced with an open vocal tract. Based on the observation of the filled pause that is gathered from the Malaysian Parliamentary Hansard Document (MPHD), the majority of filled pauses are unvaried phoneme such as /aaa/. The other two types of filled pause are /eee/ and /emm/.

The second category of Malay phoneme is consonant that is divided into seven classes. It consists of stop or plosive, affricates, nasal, glides, liquids, fricatives and semivowel. The sound from consonants is produced by air from lungs and consists of noise. The noise is generated in mouth and nose, for instance, phoneme /p/ and /b/. Figure 2.6 illustrates the consonant utterances classification for the Malay language (El-Imam & Don 2000).

The last category, miscellaneous category, consists of diphthong and vowel function. Vowel function is a combination of two different vowels (ia, io, and iu) and most often used in words absorbed directly from its English equivalent such as radio and audio, and in some original Malay words such as *nyiur* (i.e. coconut), and *hias* (i.e. decorate) (Hussain 1997). The next level of speech unit is syllable which is the combination of phonemes. For example, /ta/, /da/, /ba/ and /si/. Further observation on the MPHD shows that, the Malay language elongation is constructed from the combination of consonant and vowel, forming a syllable such as /DA/ from the word “ada”, and /TA/ from the word “kita”.

The varieties of speech sounds and types are due to different configuration of the speech organ system. For example, vibrating glottis resonates through the vocal tract produces periodic pulses of air resulting in voiced segment of the speech such as

/b/, /d/, and /g/ phonemes. Meanwhile, unvoiced phonemes such as /v/, /z/ and /h/ are produced when larynx is open and there is no vibration in vocal cords, so air flowing through the vocal tract is not periodic (Shaneh & Taheri, 2009).

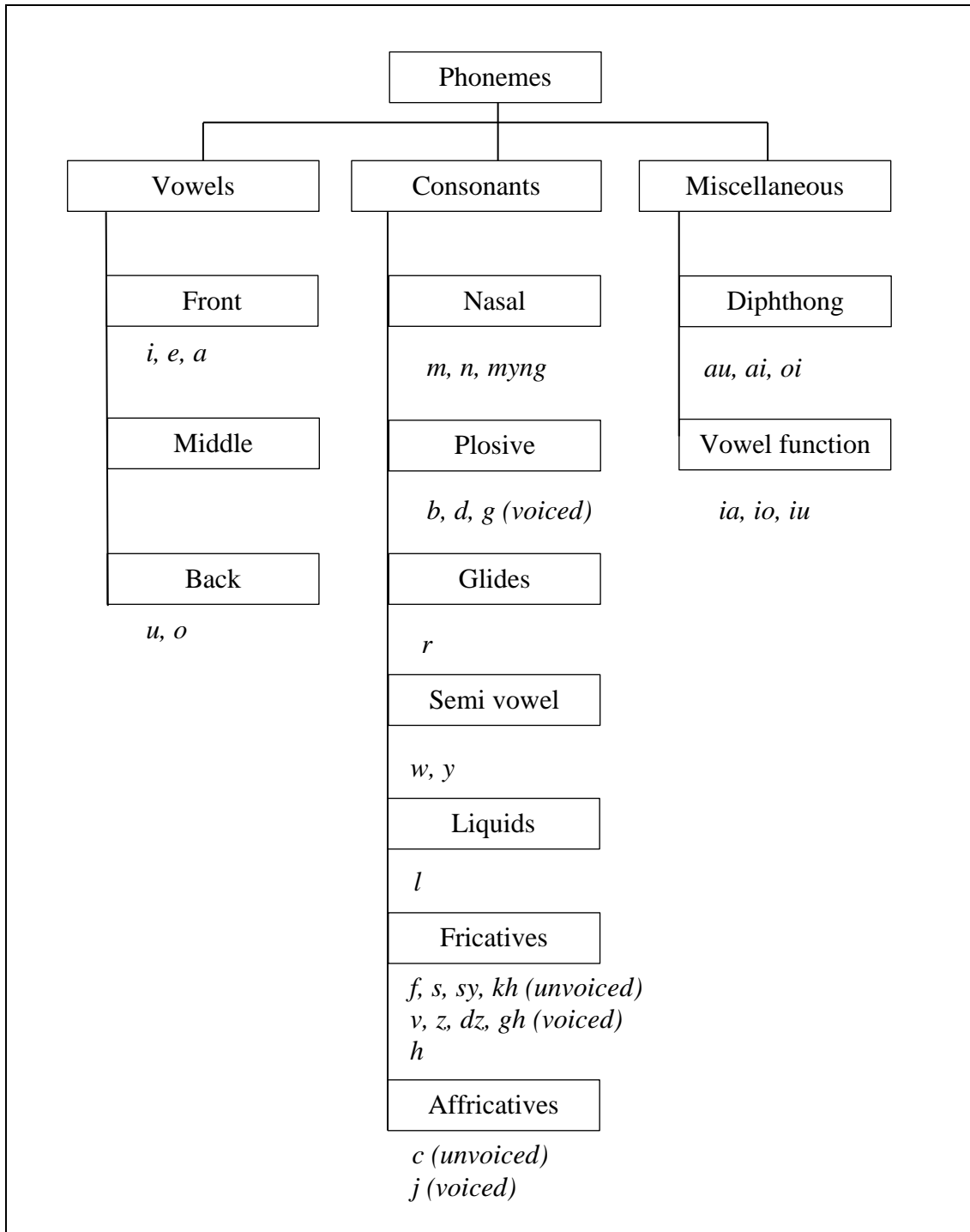


Figure 2.6: Structure of Malay Language Phonemes

Filled pause is consist of phoneme which is pronounced unvaried (Ogata et al., 2009), while elongation is vowel lengthening of a syllable or a word (Clark & Fox, 2002) containing both vowels and consonants. Since filled pause is uttered with unvaried phoneme, a periodic speech segment is produced with the utterance. Compared to elongation, the combination of consonant and vowels produced a non-periodic segment. According to Jalil et al., (2013) the narrow constriction of the vocal tract in consonant utterances obstruct a non-periodic sounds.

The main human speech production organ is the vocal tract. The overall human speech production organ is shown in Figure 2.7. One of the standard acoustical features of filled pause is stable formant frequency. Formant frequency is produced in relation with the vocal tract activity. The oral opening from the larynx to the lips and the joined of nasal and oral passage produces the vocal tract resonance. While speakers utter the filled pause, there is minimal articulation in the vocal tract that result in stable formant frequency production (Audhkhasi & Angeles, 2009).

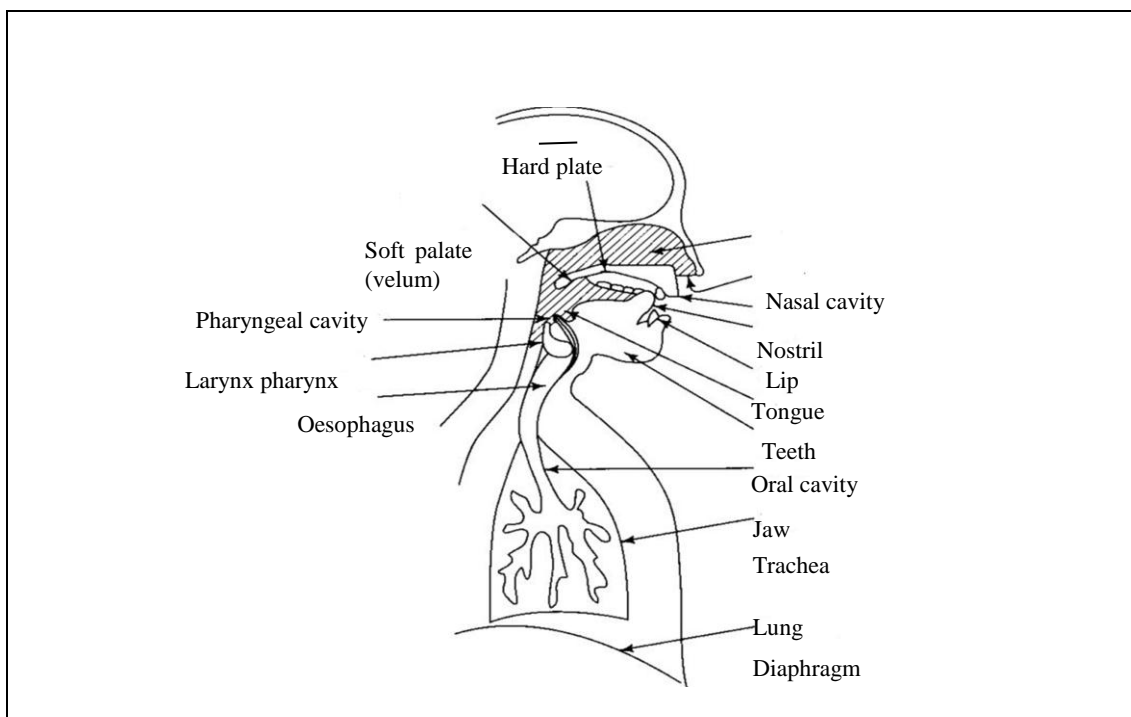


Figure 2.7: Human vocal mechanism

Filled pause is uttered based on unvaried phoneme. The utterances of unvaried phonemes produce constant energy (Jalil et al., 2013) thus constructing a periodic and continual speech signal in filled pause pronunciation. On the other hand, elongation is a combination of consonant and vowels. The production of consonant and vowel go

through different speech mechanism. According to (Fogerty & Humes, 2012) consonants are more stressed compared to vowels and more stressed compared to unstressed syllables. Consonant and vowel are two separated speech categories. Studies show that each category carries different characteristics due to different speech mechanism process (Perkell & Guenther, 2004). Generally, vowels are higher in intensity, longer in duration, lower in frequency, and slower movement in the articulators compared to consonants.

Research of filled pause in Malaysia is still at an early stage. Although there are existing filled pause studies conducted in Malaysia, majority of the researches focused on English language spoken by Malaysian or international students (Enxhi et al., 2012; Khojastehrad, 2012; Pillai, 2006). In Enxhi et al., (2012) the English language disfluencies including mispronunciation spoken by Malaysian students are studied. In their linguistic-based research, the motivation is to investigate the types of disfluencies used by students.

There was no further discussion on the acoustical features of the detected disfluencies, which is the basis of this research. Another disfluencies studies conducted in Malaysia was done by Khojastehrad, (2012) by focusing on the pattern of disfluencies among Iranian students for gender differentiation. Pillai, (2006) used English language database for the disfluencies analysis related with psycholinguistic processes of self-monitoring and self-repair.

The specific Malay language term of filled pause is also not available in the Malay dictionary even though filled pause exists in Malay conversations (Ardi, personal communication, 1st August, 2013). Up to date, only filled pause of Malay language database has been collected by Chong et al., (2012). In their research, filled pause are transcribed as ‘uh’, ‘er’ and ‘*apa ni*’ and ‘*apa tu*’ similar to “I mean” and “you know” in English language. The other disfluencies such as repetition and rephrase were also transcribed. However, further analysis on the disfluencies collection is not done and addressed.

Therefore, in this research, the Malay language filled pause and elongation word collection based on Malaysia Parliamentary Debate Session are constructed and the acoustical features are analyzed for better understanding of Malay language disfluencies.

2.9 SUMMARY

Detail study of the human vocal tract showed that the articulation of filled pause and elongations is slightly different. Filled pause is uttered with unvaried phoneme of vowels producing periodic speech segments. Meanwhile, combination of vowels and consonant in elongation creates non-periodic speech segments. Therefore, accurate representation of each types of disfluency is necessary to classify them into separate classes. Literature has shown that, well-established acoustical features are commonly used in classification of filled pause and elongation. Therefore, further investigation need to be done to identify suitable acoustical feature(s) such as ZCR, F0, FF, STE, and MFCC to model these features for optimized classification. Naïve-Bayes classification is chosen as it is a simple and efficient probabilistic classifier. Furthermore, Naïve Bayes assumes feature independence which suits the characteristics of filled pause and elongation. For the conditional probability density estimation that needs to be done in Naïve Bayes classifier, the KDE is chosen. Malay language is chosen as the research domain as reviews indicated that spontaneous speech work particularly disfluencies related studies are very much lacking. In the next chapter the methodology of this research are described.

CHAPTER THREE

METHODOLOGY

3.1 INTRODUCTION

This chapter presents the overall methodology that is implemented in this research. Other than that, the development of Malay language filled pause and elongation datasets is also elaborated as it is the first contribution of this research. From here onwards, filled pause is referred to as FP and elongation is known as ELO. The aim of this chapter is to give an overview of the overall research methodology. Overall, the methodology is divided into five stages. The flow of the methodology can be visualized as in Figure 3.1.

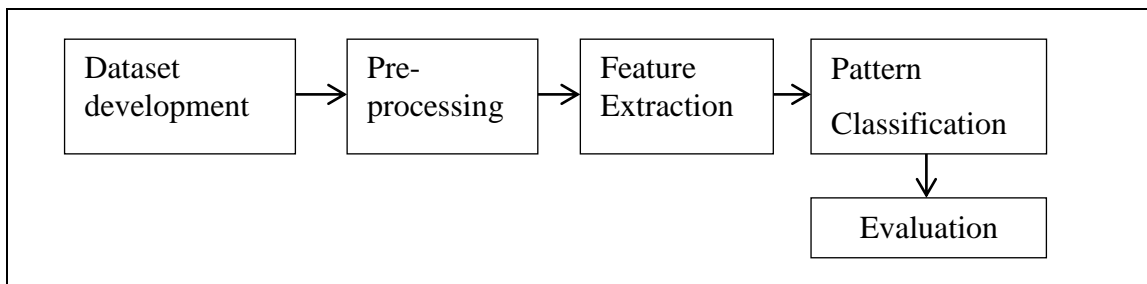


Figure 3.1: Basic flow of research methodology

The first stage is dataset development of filled pause and elongation. Filled pause dataset (i.e. FP_DATA) and elongation dataset (i.e. ELO_DATA) are then subjected to pre-processing stage which is a combination of established procedures in speech analysis. The output of the speech pre-processing is passed to the feature extraction stage process to get the feature representation of the speech. The selected acoustical feature vectors are then fed into the classification stage to classify the speech disfluencies into filled pause or elongation. The last stage is to evaluate the classifier performance based on several measurements. Detail of each block presented in Figure 3.1 is further elaborated in a research framework shown in Figure 3.2. This chapter discussed the detail processes of the data collection and analysis of Malay language disfluencies, in particular filled pause and elongation.

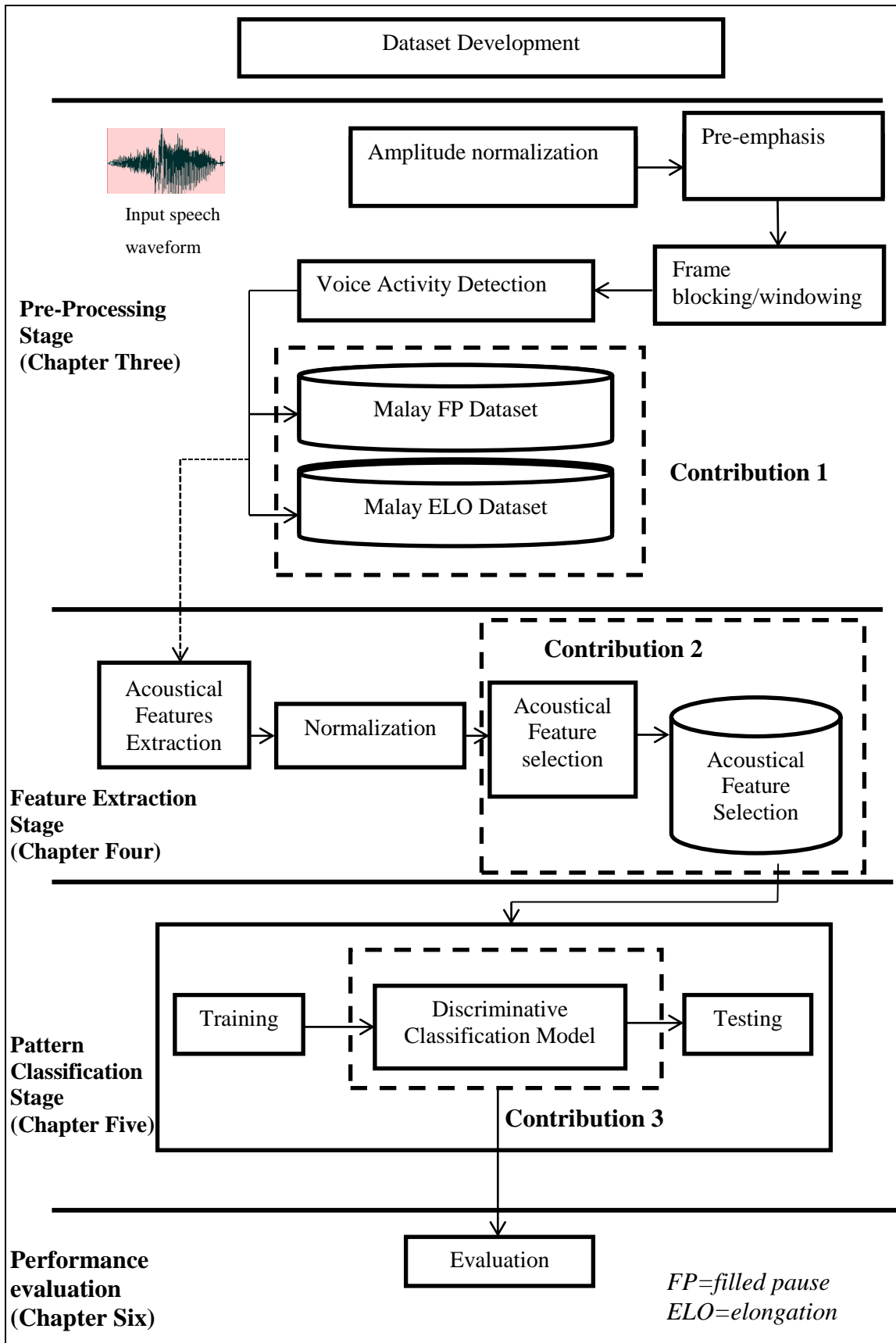


Figure 3.2: Research framework

Malay language is categorized as an under-resourced language (Besacier et al., 2013) due to the lack of resources for speech and language processing such as corpora, transcribed speech data and pronunciation dictionaries. Even though, research in Malay language is progressing, proper documentation and digital language resources are still scarce. This chapter addressed the first objective of this research that is to create Malay language filled pause and elongation datasets. The datasets are further processed and used in all experiments in this research. Basic facts relating to the Malay language and speech sounds are also presented to gain understanding of filled pause and elongation's unique characteristics.

3.2 CONSTRUCTION OF MALAY FILLED PAUSE AND ELONGATION DATASETS

In this research, speech audio is gathered from Malay Parliamentary Hansard Document (MPHD). The MPHD document contains *Dewan Rakyat* Parliamentary debate session of the year 2008. The debate session involves formal and spontaneous speeches daily recording of 222 elected members of *Dewan Rakyat*.

The speech is surrounded with medium noise condition (≥ 30 dB) which contains the presence of background noise such as electrical noise and microphone noise. Since the speech was recorded live, it also consists of speakers interruption talking in different speaking style (low, medium, high intonation or shouting). The speakers are from various ethnic background such as Malay, Chinese, and Indian. Laughter, cough and claps are other common noise elements of a live debate sessions in the MPHD. The spontaneous speech component of MPHD are uttered with various disfluencies types of filled pause, elongations, repetitions, sentence restart and correction.

The data collection process is illustrated as in Figure 3.3. In the first step, the video files of MPHD is converted to audio format by using video to audio converter freeware and named MPHD.wav. The video recording collection of MPHD comprises of 51 video files (Seman, 2012). Each video file contains a morning and an evening session that was conducted within eight to thirteen hours and is accompanied with text transcription. The analysis of video quality is done one by one to select the best perfect match between video and text transcription (Seman, 2012).

Out of 51 video files, only 22 files are suitable for further processing (Seman, 2012). They are not corrupted, no missing sounds and matched perfectly with the transcriptions (text files). These 22 audio (.wav) files contains 1 074 072 words with approximately 214 814 sentences. Only seven audio (.wav) files are chosen and analysed to extract the Malay filled pause and elongation. From preliminary observation, these seven files contain large amount of filled pause and elongation. The quantitative information analysis of the chosen files is tabulated in Table 3.1.

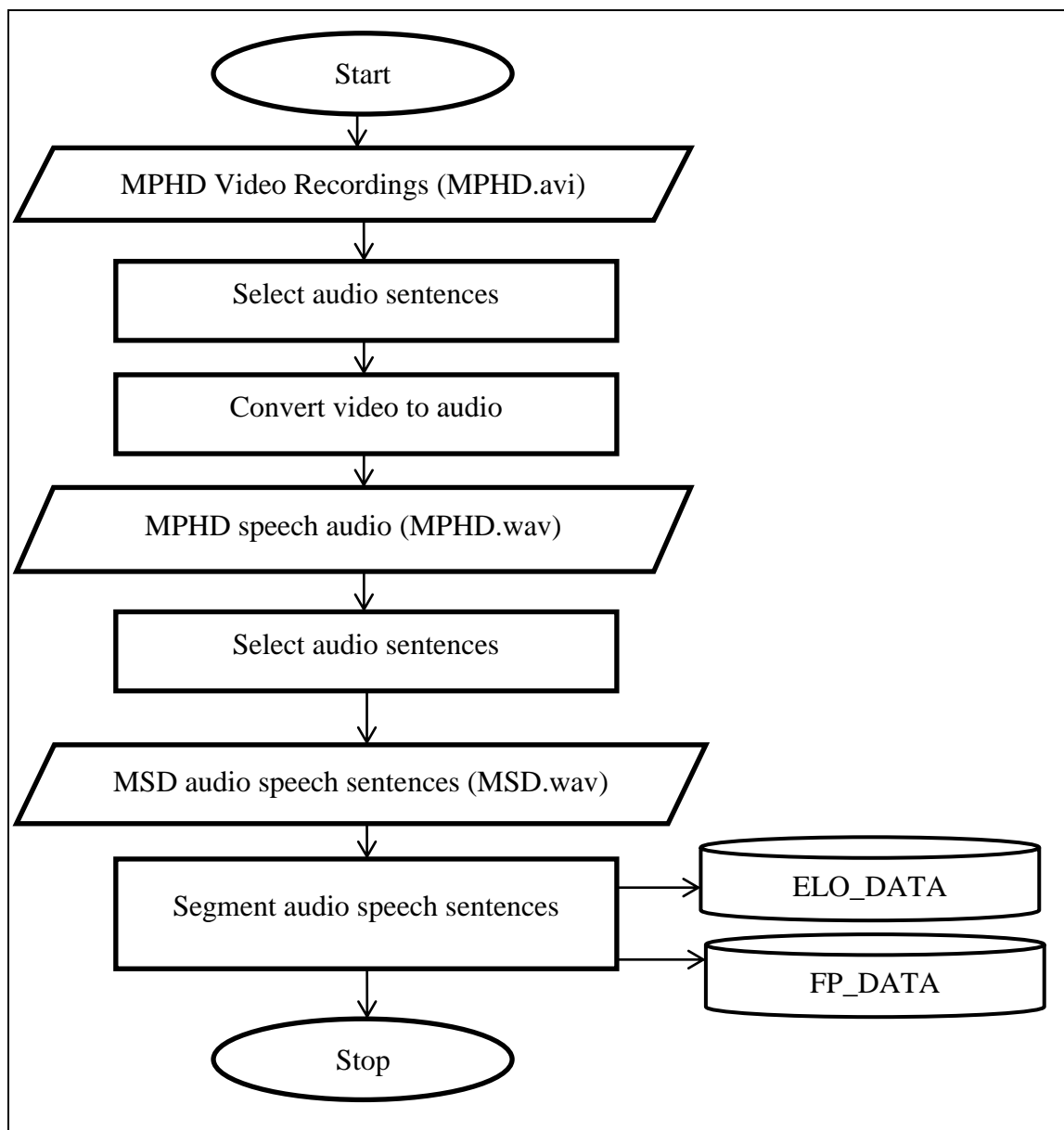


Figure 3.3: Data collection process flow

Table 3.1:
Quantitative information of selected MPHD files

No	Files	No of Topics	Duration	No of Speakers	Total Filled Pause	Total Elongation
1	DR28052008 (MEI)	11	9hrs	129	490	498
2	DR29052008 (MEI)	15	10hrs	114	300	389
3	DR07072008 (JULY)	16	13hrs	210	370	359
4	DR28082008 (AUGUST)	10	8hrs	123	600	557
5	DR10112008 (NOVEMBER)	10	8hrs	105	500	450
6	DR03112008 (DECEMBER)	12	13hrs	152	420	397
7	DR11122008 (DECEMBER)	10	8hrs	143	320	350
Total		84	69hrs		3000	3000

The seven (.wav) selected audio files as listed in Table 3.1 are resampled at 16kHz sampling rate and quantized at 16 bits per sample. According to the sampling theorem, frequency sampling is described as in equation 3.1.

$$fs = 2fc \quad (3.1)$$

where,

fs = frequency sampling

fc = the highest frequency contained in the signal

The process of resampling is done by using Audacity version 2.0.6; a freeware tools for speech analysis. Based on the observation of the sentences, the disfluencies gathered from MPHD's of Malay spontaneous speech are classified into three categories, which are filled pause, elongation, sentence restart and repetition.

From the seven chosen files, 1348 sentences are chosen manually by listening to the occurrences of filled pause and elongation in the MPHD audio. Each sentence is selected based on the occurrence of filled pause or/and elongation. These sentences construct the first dataset and named as Malay Sentence Data (MSD) that is used to obtain filled pause and elongation. Based on observations of the MSD, some speakers only used filled pauses to maintain the conversation while others used both filled pause and elongation. Therefore, 148 sentences of MSD consist of sentences with filled pause and normal words only (Figure 3.4) and the remaining 1200 sentences contain filled pause, elongation and normal words (Figure 3.5). The examples of sentences that contain filled pause, normal words and elongation are presented in Figure 3.4 and Figure 3.5. In the figures, the filled pause is marked using dashed-oval shape while normal word is marked in dashed-rectangle and the elongation is marked

in dashed-square shape. The silence is transcribed as *sil* in the transcription pane above the speech waveform.

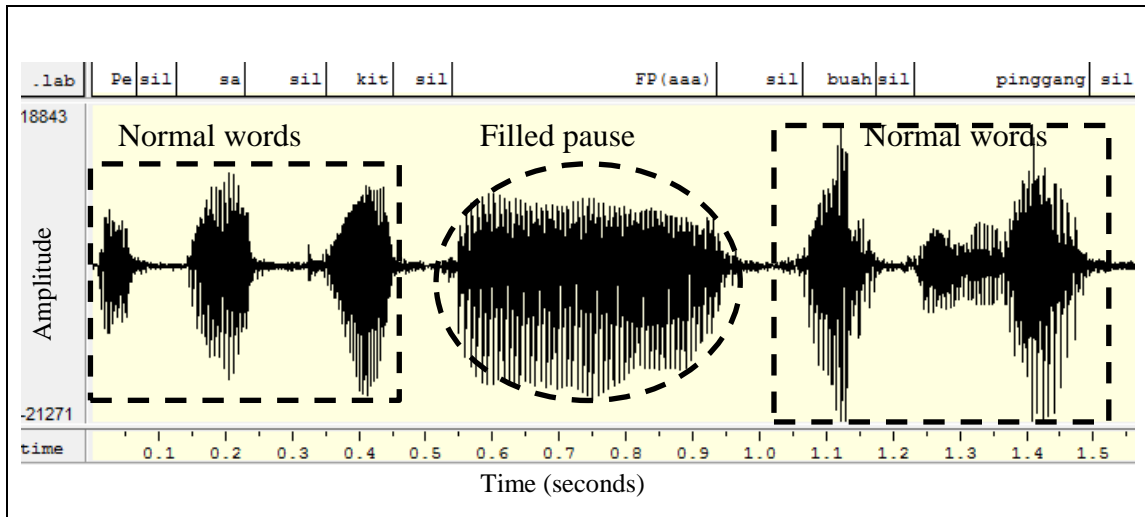


Figure 3.4: A complete sentence with occurrences of filled pause only (Malay sentence id S169M9T04: *Pesakit aaa, buah pinggang*)

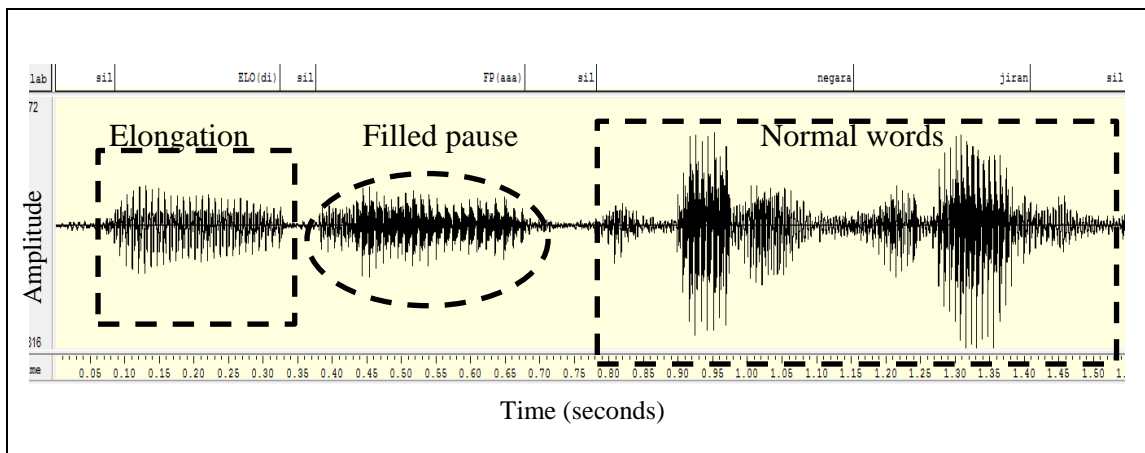


Figure 3.5: A complete sentence with occurrence of filled pause, normal words and elongation (Malay sentence id S53M5T03: *di(ELO) aaa(FP) negara jiran*)

The description of each segmented sentences is given by following the notation symbol “S” (number of sentence), “F/M” (gender) “T” (topic number) and the segmented isolated filled pause and elongation is based on the number of sentence followed by number of filled pause. For example, the sentence in Figure 3.5 is labeled as S53M5T03 with the corresponding filled pause and elongation of the sentence is F53 and E53. Subsequently, in order to gather different sets of filled pause and elongation data collection, all sentences are manually segmented for further used in this research. A total of 3000 isolated filled pause comprising 2400 ‘aaa’, 450 ‘eee’ and 150 ‘emm’ is collected from the segmented sentences (i.e. MSD dataset) and

named as FP_DATA. Meanwhile, 3000 elongations are also extracted from MSD dataset and called ELO_DATA. In order to get an accurate endpoints segment, voice activity detection (VAD) techniques will be applied in both datasets (FP_DATA and ELO_DATA). Furthermore, the datasets have been verified by the linguist experts to confirm that the collection only contains the filled pause and elongation of word segments. (*Refer to Letter of Approval in Appendix B*).

3.3 MALAY LANGUAGE FILLED PAUSE

In this section, further discussion of filled pauses used in the MSD database is presented. Upon closer inspection of the speech data done by the linguist experts (Norizah Ardi, Academy of Language Studies Universiti Teknologi MARA), Malay language filled pause can be labeled into three categories which are /aaa/, /eee,/ and /emm/. Figure 3.6, Figure 3.7, and Figure 3.8 illustrate each type of the filled pause graphically.

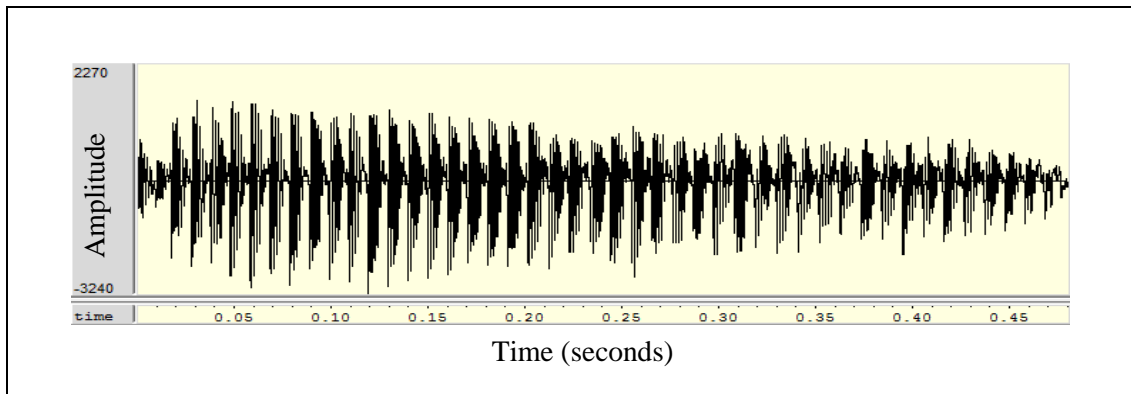


Figure 3.6: Filled pause /aaa.wav/

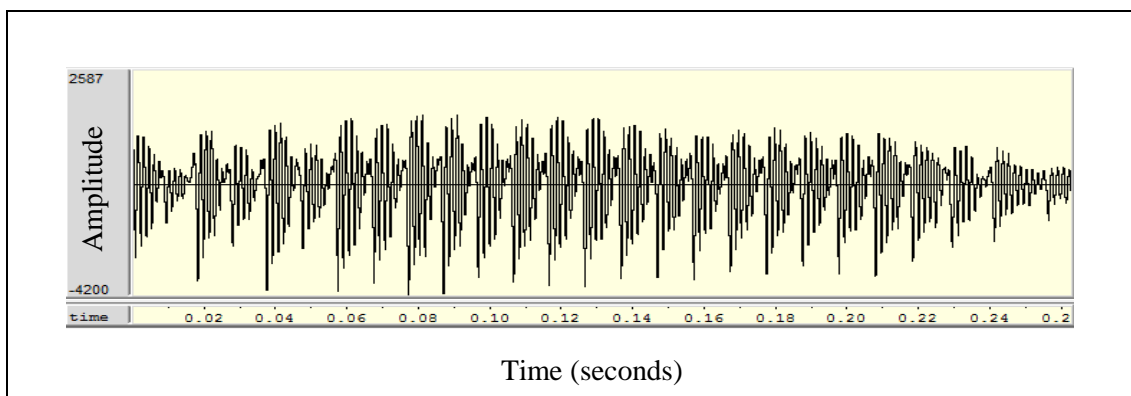


Figure 3.7: Filled pause /eee.wav/

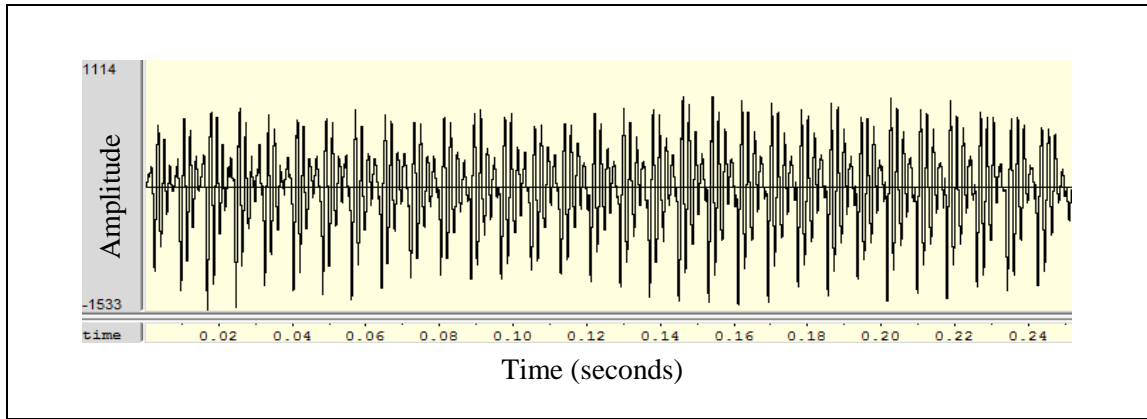


Figure 3.8: Filled pause /emm.wav/

In this research, filled pause are gathered from speech segments that are longer than 200ms and those that fit into filled pause type /aaa/, /eee/ or /emm/. Another considered characteristic is all filled paused selected have adjacent silent.

3.3.1 Established Filled Pause Acoustical Characteristics

As stated in Chapter Two, filled pause is represented using standard and well-established acoustical features such as flat fundamental frequency and stable formant frequency compared to normal words. The flatness and stableness of the fundamental and formant frequency are measured by observing the standard deviation of each feature.

The lower the standard deviation indicates a more stable speech segment, thus filled pauses are supposed to have lower features' standard deviation compared to normal words. Both features of formant and fundamental frequencies are calculated by using autocorrelation and linear prediction coding, respectively. The detail process of extracting fundamental and formant frequency are presented in Chapter Four. A preliminary analysis of the acoustical features of filled pause and normal words is done and the results are tabulated in Table 3.2. The objective of this analysis is to compare characteristics of filled pause and normal words using their acoustical features. One hundred samples of filled pause from FP_DATA and 100 normal words from MSD each are selected and the average standard deviation of F0 and FF are calculated. It can be seen that the average standard deviation of F0 for filled pause is much lower than normal words. This confirms that filled pause's F0 is flatter than normal words. Formant frequencies of filled pause and normal words are observed at

four levels (i.e. FF1, FF2, FF3, and FF4). Even though the average standard deviation varies between each level, filled pause consistently shows lower standard deviation than normal words at all levels. Again, this indicates that formant frequency of filled pause is more stable than normal words.

Table 3.2:

Average standard deviations of acoustical features measurements of filled pause and normal words

Feature	Filled pause	Normal words	Differences
Fundamental frequency	47	120	73
Formant Frequency1	38	132	94
Formant Frequency2	148	389	241
Formant Frequency3	194	513	319
Formant Frequency4	175	310	135

According to Audhkhasi & Angeles, (2009), there are two types of filled pause in spontaneous speech that are robust and non-robust. Robust filled pauses are filled pauses with at least one of the characteristics as described below:

i. Longer duration

The optimal duration of robust filled pause is $>200ms$ as previously done in Gaurav & Nigel, (2006); Li et al., (2008); Stouten, (2008).

ii. Stable spectral energy

The spectral energy rises quickly at the start and remains stable in the middle, then falls gradually at the end of the filled pause.

iii. Adjacent silent.

Robust filled pause comprises silence in adjacent segments i.e. before and after the filled pause.

In contrast with robust filled pause, characteristics of non-robust filled pause are the existence of intonation contour, embedded within normal words and short duration. Figure 3.9 until Figure 3.11 show the example of robust and non-robust filled pause, respectively. It can be seen in Figure 3.9 that the robust filled pause has adjacent silence, longer duration and stable spectral energy. Whereas, filled pause in Figure 3.10 is short in duration and embedded within normal word segments. The non-robust filled pause characteristic can be observed in Figure 3.11, which shows that it has expressive intonation contour. In this research, filled pause are gathered from speech segments that are longer than $200ms$ and those that fit into filled pause type /aaa/, /eee/

or /emm/. Another considered characteristic is all filled paused selected have adjacent silent.

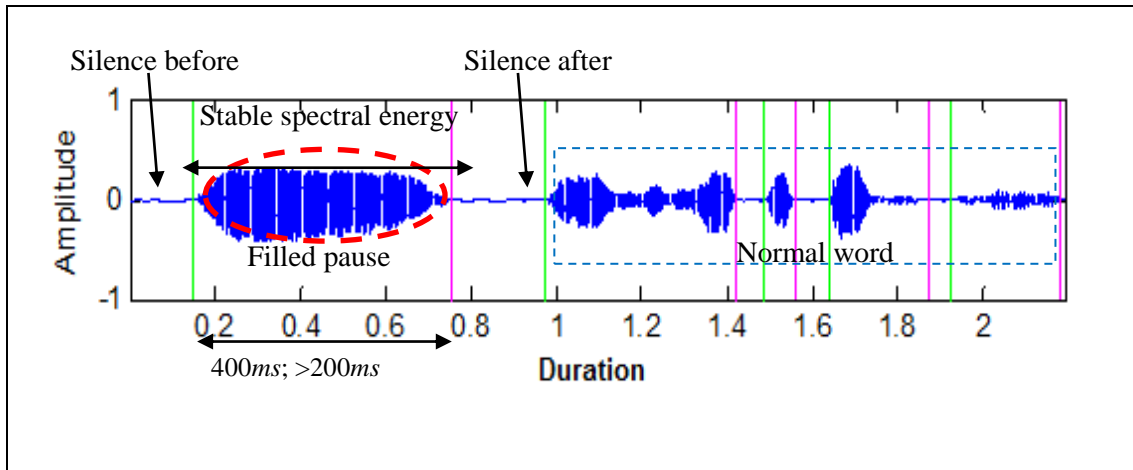


Figure 3.9: Robust filled pause

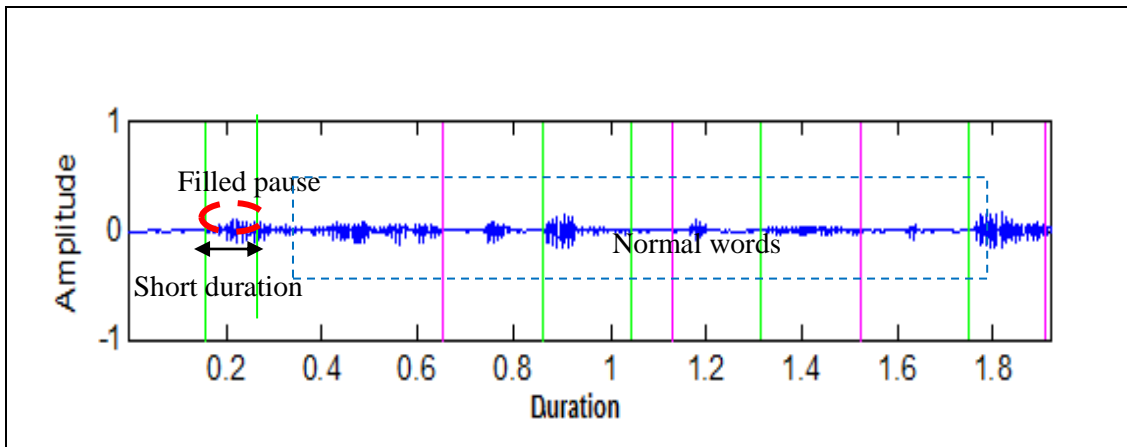


Figure 3.10: Non robust filled pause with short duration and having word insertion

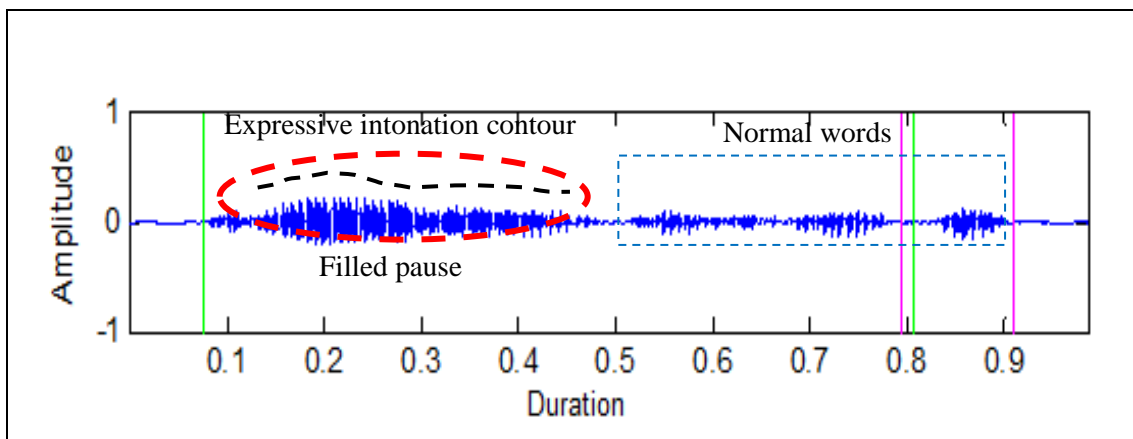


Figure 3.11: Non robust filled pause with expressive intonation contour

3.4 MALAY LANGUAGE ELONGATION

The second class of speech disfluencies that is studied is elongation (ELO). Elongation appears in the form of syllable such as /da/ and /ta/ from the word /Ada/ and /kita/ respectively. Based on the literature, Mandarin language elongation can occur at any position in the utterance; whether at the start, in the middle, or at the last syllable of a word (Lee et al., 2004). While English elongation can appear at the first or the last syllable (Gaurav & Nigel, 2006).

Malay language is different from English language because Malay words are agglutinative alphabetic-syllabic based on four distinct syllable structures, i.e. V, VC, CV and CVC. The agglutinative nature of Malay language ensures the combination of consonant and vowel in elongation causing non-periodic speech segments. To gain a better understanding of Malay language elongations, 100 elongations are randomly observed from each file in MPHD datasets. In this research, the elongations are extracted from the normal words which syllables are more than 200ms. Elongations may occur in monosyllabic words (e.g. /nya/, /ke/, /di/) or words with multi syllables (e.g. /kita/, /negara/, /mereka/). However, majority of the elongation is produced by speakers at the end of the word syllable for example the word /ada/. In the word /ada/, the syllable /da/ is being extended producing elongation. An example of elongated word is shown as in Figure 3.12.

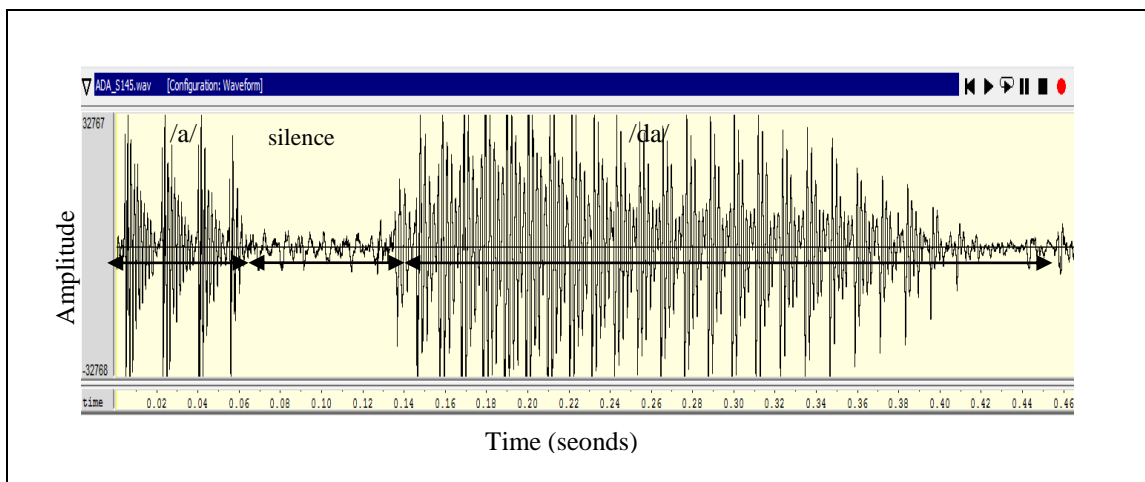


Figure 3.12: Elongation word at the last syllable of /da/ from word /a//da/

The common elongated words that are extracted from MPHD data are illustrated in Table 3.3. It can be seen that, the last elongated syllable always ended with a vowel.

Table 3.3:
Common Elongated Word in the MPHD Data

Word (type)	English translation	Structure	Word	English Translation	Structure
Ada	have	V+CV	Maka	Therefore	CV+ CV
Bahawa	that	CV+CV+CV	Mereka	They	CV+CV+CV
Beberapa	How many	CV+CV+CV+CV	Negara	Country	CV+CV+CV
Bila	when	CV+CV	Nya	Him/Her/It	CCV
Cuma	Only	CV+CV	Pada	To	CV+CV
Tua	Old	C+VV	Paksa	Force	CVC+CV
Dua	Two	C+VV	Pertama	First	CVC+CV+CV
Harga	Price	CVC+CV	Peserta	Contestant	CV+CVC+CV
Juga	Also	CV+ CV	Saya	I	CV+CV
Kata	Speak	CV+ CV	Secara	Way	CV+CV+CV
Kerana	Because	CV+CV+CV	Tanya	Ask	CV+CCV
Kira	Count	CV+ CV	Warna	Color	CVC+CV
Itu	That	V+CV	Tapi	But	CV+CV
Ini	This	V+CV	Satu	One	CV+CV
Jadi	So	CV+CV	Tunggu	Wait	CVCC+CV
Di	At	CV	Ke	To	CV

3.5 PRE-PROCESSING

All the speech data that are used in this research are pre-processed for the purpose of feature extraction. In the pre-processing stage, several processes are undertaken inclusive of amplitude normalization, pre-emphasis, framing and windowing and voice activity detection. The preprocessing of speech is a vital stage in any speech processing research (Keerio et al., 2009). Each of the pre-processing process is discussed in the following subsections.

3.5.1 Amplitude Normalization

The raw speech data are a collection of speech uttered by different speakers thus the amplitude and energy vary. The variety of speaker's speech energy can cause error or unstable classification rate if the feature vector is directly extracted. Therefore, the purpose of amplitude normalization is to ensure that the level of the energy is standardized or similarly calibrated. In this research, the z-score

normalization technique is adopted. The speech amplitude variability is normalized to have zero mean and one standard deviation. Speakers' volume variations need to be normalized before the next process is taken so that the volume will not become a performance degradation factor. The steps of speech vector normalization are illustrated as in Figure 3.13.

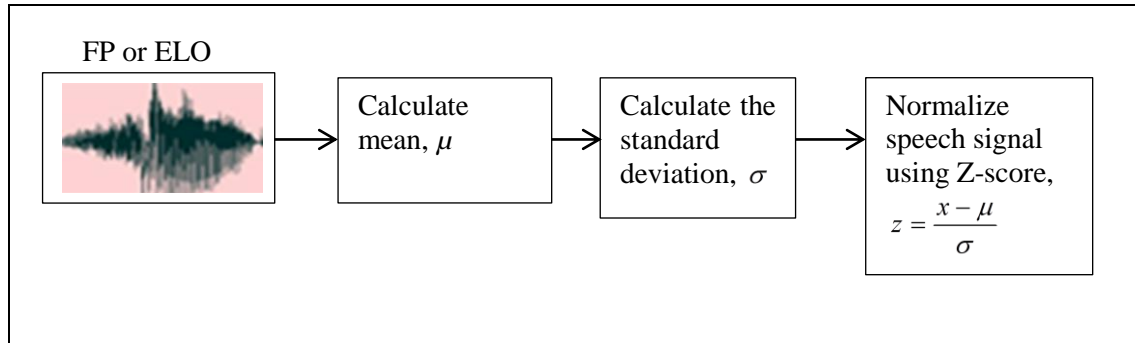


Figure 3.13: Steps of amplitude normalization on the speech signal

The normalization steps are as follows:

- i. The mean, $\mu(x)$ of the speech vector is computed
- ii. The standard deviation, $\sigma(x)$ of the speech vectors (x) is computed
- iii. By using the mean and standard deviation calculated in step (i) and step (ii), the normalized speech vector $z(x)$ is computed as in equation (3.2).

$$z(x) = \frac{x - \mu(x)}{\sigma(x)} \quad (3.2)$$

where

x = speech vector

The normalization effect is evaluated by calculating the mean amplitudes of the speech samples (3000 FP and 3000 ELO). The mean amplitudes variance before and after the amplitude normalization are compared and shown in Table 3.4. From the result, it is a clear evident that the mean amplitude variance after the normalization is smaller compared to before speech vectors normalization. Smaller variance shows that the difference between normalized amplitude among the filled pauses and elongations is very minimal. As stated earlier, the amplitude normalization is important to ensure the energy of the speeches is within the same range.

Table 3.4:
Mean amplitude variance due to normalization

Before		After	
FP	ELO	FP	ELO
3.8066×10^{-6}	3.7886e-06	2.1778×10^{-32}	2.1587×10^{-32}

The output of normalized speech signal, $z(x)$ is used as input to proceed with the pre-emphasis stage.

3.5.2 Pre-Emphasis

Generally, digitized speech waveforms comprise additive noise and have high spectral dynamic range. For example, a low energy can be found in high frequency spectrum of a speech as well as high energy in low frequency spectrum. Because of that reason, a process called as pre-emphasis is performed on the normalized speech $z(x)$ to flatten the speech spectrum and to emphasize the high-frequency part of the speech signal that was repressed through the human sound production mechanism. For example, pronunciation of vowels existing in filled pause and elongations have high energy (Kitamaya et al., 2003) and may be pronounced at the lower frequency. Therefore, it needs to be boosted to attenuate the information from the higher frequency for better acoustical feature representation. The most extensively used pre-emphasis digital high-pass filter is defined as in equation (3.3).

$$y(n) = z(x) - A \times (n-1) \quad (3.3)$$

where:

- $y(n)$ = the value of output signal at discrete time step n
- $z(x)$ = the value of normalized input signal at discrete time step n
- A = is a constant normally set between 0.9 to 1

In this research, the value of 0.95 is chosen as A . In the literature, there are various usages of pre-emphasis constant. Yusof et al (2007) used a constant of 0.95 for pre-emphasis process. While in Meseguer, (2009), the pre-emphasis constant is set to 0.97. However, according to Abbas, (2013), the typical value of pre-emphasis constant is 0.95. A low frequency signal is the one with slow time variation. The slow variation

effect on low frequency signal concurrently produces adjacent samples of similar numerical value. From equation (3.3), the subtraction process removed the part of the samples that did not change in relation to its adjacent samples to retain the high-frequency components. The output signal of the pre-emphasis process $prem^x(n)$ is then past to the framing stage.

3.5.3 Framing

Speech signal is non-stationary and non-periodic in a longer duration. Its statistical properties are non-constant over time. However, practically, a speech at a frame of $20ms \sim 30ms$ is considered stationary and quasi-periodic (Ganapathy, 2012). Thus, the non-stationary properties of a speech signal need to be transformed as stationary using framing. Framing a speech signal is a process of blocking the speech signal into frames of N samples, with adjacent frames being separated by M samples i.e., the frame is shifted with M samples from the adjacent frame. The spectral features estimated from frame to frame will be smooth if the shifting is small. The shifting process is important to ensure overlapping of the speech frame. The absence of overlapping between adjacent frames will cause the speech signal to be entirely mislaid and will contain noisy components only.

The general equation for frame blocking is written in equation (3.4) by assuming that the speech frame length (l_{th}) is represented as S and the entire speech signal is denoted as L .

$$X_l(N) = \bar{S}(M_l + N) \quad (3.4)$$

where.

- X_l = frame of speech
- N = $0, 1, \dots, N-1$ sample
- l = $0, 1, \dots, L-1$ frames

In this research, the frame size is set to $20ms$ (320 points) frames, which are overlapped at $10ms$ (160 points). A typical frame shift of $10ms$ of a short frame of $20ms$ is always chosen in speech processing research (Singh et al., 2012). The overlapping is important to ensure the smooth transition of estimated parameters between frames.

3.5.4 Windowing

Windowing is done to reduce the discontinuities of the speech signal at the edges of each frame by applying a tapered window to each frame. At each framed speech signal, a window is applied at the beginning and ending by using window function. For a window $w(n)$, the windowed signal will be defined as in equation (3.5).

$$\bar{y}(n) = x(n).w(n), 0 \leq n \leq N - 1 \quad (3.5)$$

where,

$w(n)$ = Hamming window

$x(n)$ = speech signal

$\bar{y}(n)$ = windowing result of the signal

Hamming window is the mostly used windowing function applied on each speech's frame of the speech and is described in equation (3.6). Hamming is chosen because it provides better frequency resolution as it minimizes signal discontinuity (Shreve, 1995).

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right) & , \quad 0 \leq n \leq N \\ 0 & \end{cases} \quad (3.6)$$

3.5.5 Voice Activity Detection

In speech processing research, the utterance consisting of speech, silence and other background noises need to be processed (Singh et al., 2012) . The detection of the speech presence embedded in various types of unvoiced events and background noise is called end point detection, speech detection or voice activity detection (VAD). In this research, the term VAD is used. The aim of this phase is to remove the noise and silence region and to locate the exact start and endpoint of the speech utterances. In the previous data collection process, the noise and silence segments may be included in the filled pause and elongation segments and this affect the accuracy of the feature extraction. The noise and silence could be a result of the delay before a

speaker pronounces the word. It can also happen due to the gradual stop of the audio recorder during recording of the voice sample. There are various methods to perform the task of VAD. The methods include energy thresholding, pitch detection, spectrum analysis, zero crossing rate, periodicity measure, hybrid detection and fusion (Mohamad, 2009). The conventional method introduced by Rabiner & Sambur (1975) that is widely used is the combination of threshold gathered by energy and Zero Crossing Rates (ZCR) computation. Before proceeding to any other methods of VAD, energy and ZCR is discussed first as they are the basic methods of VAD. For easier understanding, this subsection is divided into two categories which is energy and zero crossing rates.

3.5.5.1 Energy

Energy is the measure of loudness of the sounds that can be used as a feature to differentiate between voiced and unvoiced signal. Vowels such as *a*, *e*, *o*, *i*, *u* are voiced signals that are higher in energy compared to unvoiced signals such as *s*, *f*, and *h*. However, the energy of silence region in speech interval is lower than the unvoiced speech signals. The energy of a speech segments can be viewed in two terms which is long and short (Sakhnov, Verteletskaya, & Simak, 2009). The long term energy of a speech is shown as in equation (3.7) and (3.8).

$$E = \sum_{m=-\infty}^{\infty} x^2(m) \quad (3.7)$$

$$E_n = \sum_{m=n-N+1}^n x^2(m) = x^2(n-N+1) + \dots + x^2(n) \quad (3.8)$$

where,

E = energy of the speech signal

$x(m)$ = speech signal

N = frame length

n = $0, 1T, 2T \dots N$

T = frame shift

In the above equations, E represents energy of the signal $x(m)$. However, speech is a time varying signal. Thus, this equation is not effective in speech energy measurement. To take advantage of the energy features of the speech, the speech needs to be chunked into smaller segments as discussed in *Section 3.5.3*. The energy is calculated based on each framed and windowed segment.

The short term energy can be defined as in equation (3.9).

$$E_n = \sum_{m=n-N+1}^n [x(n)w(n-m)]^2 \quad (3.9)$$

The absolute short term energy (ASTE) is shown in equation (3.10).

$$E_n = \sum_{m=n-N+1}^n |x(n)w(n-m)| \quad (3.10)$$

where,

$w(n-m)$ = window

n = the sample that the analysis window is centered on

N = frame length

This energy-based VAD is suitable for speech with at least 30db signal to noise ratio. The signal to noise ratio of the MPHD data used in this research is higher than 30db (Izzad et al., 2013). Initially, the mean and standard deviation of the ASTE value is calculated from the first 100ms of the speech. The first 100ms of the speech is considered silences. Then, the calculated mean and standard deviation information is utilized to find the peak energy (PE) of the entire speech sample and the silence energy (SE). Subsequently, the PE and SE are used to set two thresholds; upper T_U and lower T_L , respectively. The T_L and T_U is computed as in equation (3.13) and (3.14).

$$I_1 = 0.03 \times (PE - SE) + SE \quad (3.11)$$

$$I_2 = 4 \times SE \quad (3.12)$$

$$T_L = \text{MIN}(I_1, I_2) \quad (3.13)$$

$$T_U = 5 \times T_L \quad (3.14)$$

In equation (3.11), I_1 is a level which is three percent (3%) of the peak energy, whereas I_2 in equation (3.12) is a level, set at four times the silence energy. The lower threshold, T_L in equation (3.13), is the minimum energy between I_1 and I_2 . The upper threshold T_U , in equation (3.14), is five times larger than the lower threshold (Rabiner & Sambur, 1975). The voice activity detection process of each disfluencies (FP and ELO) is stated as below:

- i. The algorithm searches the beginning of the interval by searching two points in the speech sample where both T_L and T_U are exceeded. It assumes that the beginning of the point lies outside this interval. These two points are marked as preliminary start point (N_1).
- ii. The algorithm then searches the point at which the energy point falls below the upper threshold T_U . This point is then declared as preliminary endpoint (N_2). The examples of T_L and T_U locations and the start N_1 and end point N_2 of the utterance are illustrated in Figure 3.14. The horizontal red and green line is the T_U and T_L in the first subplot, and the vertical red and green line is the N_1 and N_2 point in the second subplot.

However, using the upper and lower threshold set by the energy based feature, the algorithm failed to detect the start or endpoint accurately for some cases. In this research, energy-based VAD often fail to detect accurately an elongated syllable with low energy speech segments such as the word /paksa/, /kira/, and /saya/ Therefore, another threshold using ZCR is used to be combined with the energy-based VAD.

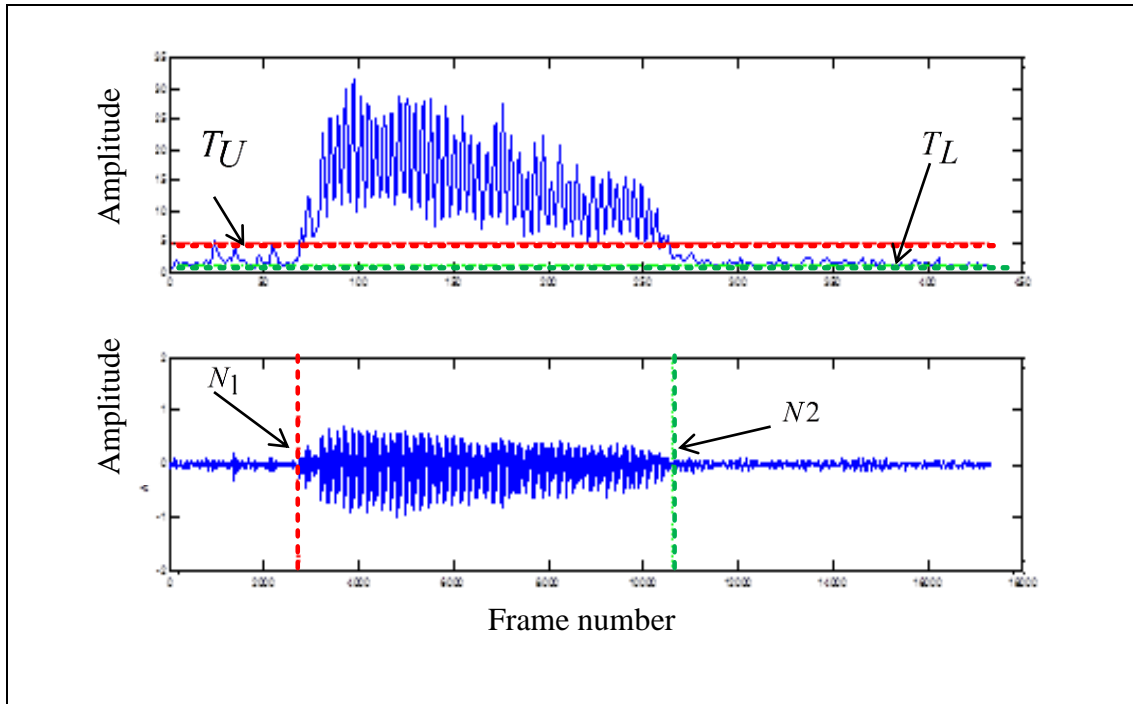


Figure 3.14: Location of upper and lower energy threshold

3.5.5.2 Zero crossing rates

ZCR functions as the second threshold to detect the low-energy phonemes that usually occur at the start and end-point of the speech boundaries. In this research, the rate of zero (level) ($Z(l)$) is measured based on the counts of the $Z(l)$ crossing per 20ms interval and overlapping at 10ms as in equation (3.15). As stated earlier in Section 3.7.3, the framing process partitions the speech signal into 20ms frame with 10ms overlap.

$$Z(l) = \frac{1}{N} \sum_{n=m-N+1}^m \left| \frac{\text{sgn}[x(n+m)] - \text{sgn}[x(n+m)-1]}{2} \right| \quad (3.15)$$

where

$$\text{sgn}[s(n)] = \begin{cases} +1, & x(m) \leq 0 \\ -1, & x(m) < 0 \end{cases}$$

Z = zero crossing rate

m = speech samples of overlapping at 10ms

n = 0, 1, ..., N-1 (N speech sample)

N = frame length

Theoretically, the zero crossing definition is “*the number of times in a sound sample that the amplitude of the sound wave changes sign*” (Bachu et al., 2008). In clean speech, the zero crossing counts of silence region are zero (Greenwood & Kinghorn, 1999). According to (Atal & Rabiner, 1976), the zero crossing rates for voiced segments in a speech is lower (i.e. at a count of 0-30) compared to unvoiced segments (i.e. 10-100 counts). This is because an unvoiced segment is produced due to excitation of the vocal tract by a noise-like source tightened by a point in the vocal tract inner. For silence segment, the zero crossing counts are still lower than the unvoiced segment and relatively similar with the voiced part of the speech. The zero crossing threshold T_{ZC} is calculated as in equation (3.16).

$$T_{ZC} = \text{MIN}(IF, \overline{ZCR} + 2\sigma\overline{ZCR}) \quad (3.16)$$

The steps of ZCR threshold calculation are:

- i. Calculation of the mean for ZCR, \overline{ZCR}
- ii. Calculation of the standard deviation for ZCR, $\sigma\overline{ZCR}$
- iii. Summation of the \overline{ZCR} and $2\sigma\overline{ZCR}$

IF is a fixed threshold with 25 crossings per 20ms (Rabiner & Sambur, 1975).

The searching algorithm of the VAD starts to pursue the stiff endpoint by moving backward from N_1 and forward from N_2 and making comparisons with the ZCR to the T_{ZC} . If the zero crossing rates exceeded the threshold by three or more times, the beginning point N_1 will move back to the first point at which the zero crossing thresholds exceeded $\overline{N1}$ (Figure 3.15). Otherwise, N_1 is defined as the beginning of the point. A similar procedure is followed at the end of utterance to remove the silence samples.

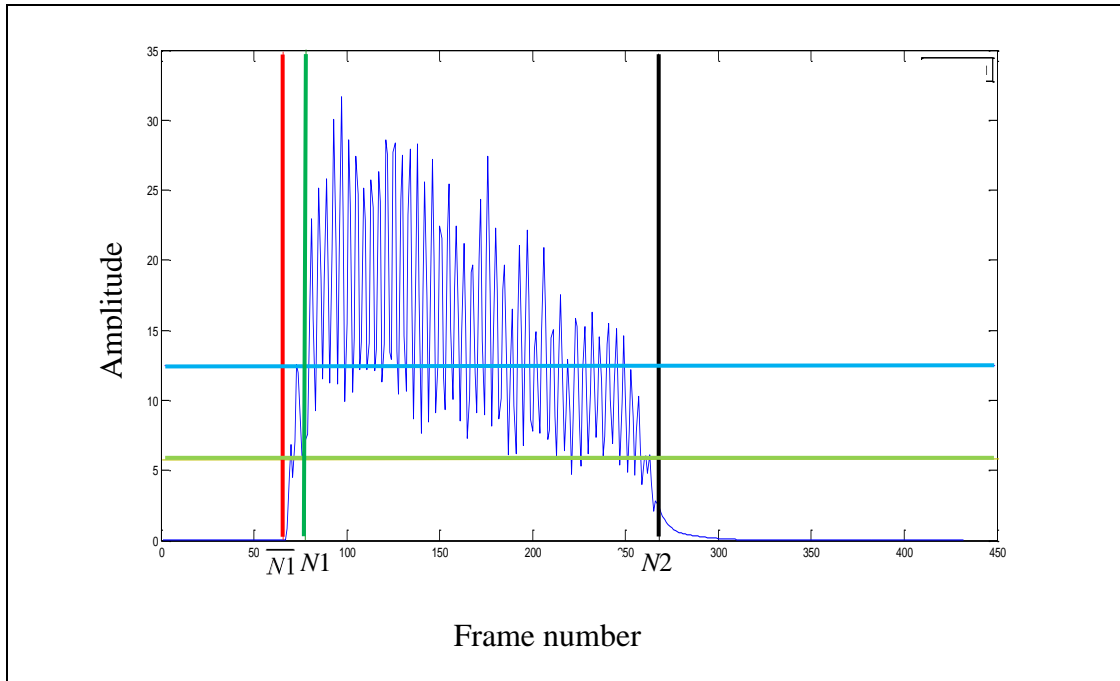


Figure 3.15: Location of $N1$ and $N2$ level

3.5.5.3 Energy and higher-order differences

The next approach that is implemented in the VAD stage is by combining the energy with high-order differences (HOD) of the speech. In this approach, the time domain characteristics of high-order differences of a given signal are utilized. A random increasing number of order that starts with the default order of 1 is used to show the impact on VAD on the speech. The statistical information that is gathered from the energy and higher order differences VAD method is presented in Table 3.5. From the table, it can be seen that there are two numbers of start point detection (1088 and 1216) and three different end point detections (5312, 4928 and 5184). These points indicate the i^{th} number of sample of the speech.

Table 3.5:
Statistical information of the HOD method of VAD

Number of Order	Start point detected	End point detected
1	1216	5312
20	1088	5312
30	1088	5312
50	1088	4928
100	1216	5184
150	1216	5312
170	1216	5312

After comparing the effects of different level of order on VAD, it can be seen from Figure 3.16 (A) that the first order differences (FOD) is suitable to be used. From the figure, it is clearly observed that the voiced region is detected accurately (refer to green and purple lines). An example of 50th order is taken and shown in Figure 3.16 (C). It is observable that the higher order differences of 20, 30, 50 and 100 have lost some information as shown by the green line in Figure 3.16 (D). It is also noticeable in Figure 3.16 (C), that wrong detection of the starting and endpoint resulted in inclusion of more unvoiced region and exclusion of more voiced region. The subsequent steps of the HOD algorithm are as follows:

- i. The energy (E) is calculated as the same process in *Section 3.5.5.1*
- ii. The absolute value of the sum of the I^{st} order differences (H) volume weight, vw is chosen from [0-1] to get the value of E and H which is used to compute the threshold T_{HOD} as in equation (3.17).

$$EH = vw \times E + (1 - vw) \times H \quad (3.17)$$

- iii. The threshold T_{HOD} is computed using equation (3.18).

$$T_{HOD} = EH \text{ min} + (EH \text{ max} - EH \text{ min}) \times r \quad (3.18)$$

where

$$EH \text{ min} = \frac{N \times k}{100}$$

$$EH \text{ max} = N - EH \text{ min} + 1$$

N is the frame number and k is the parameter setting for y-axis percentile.

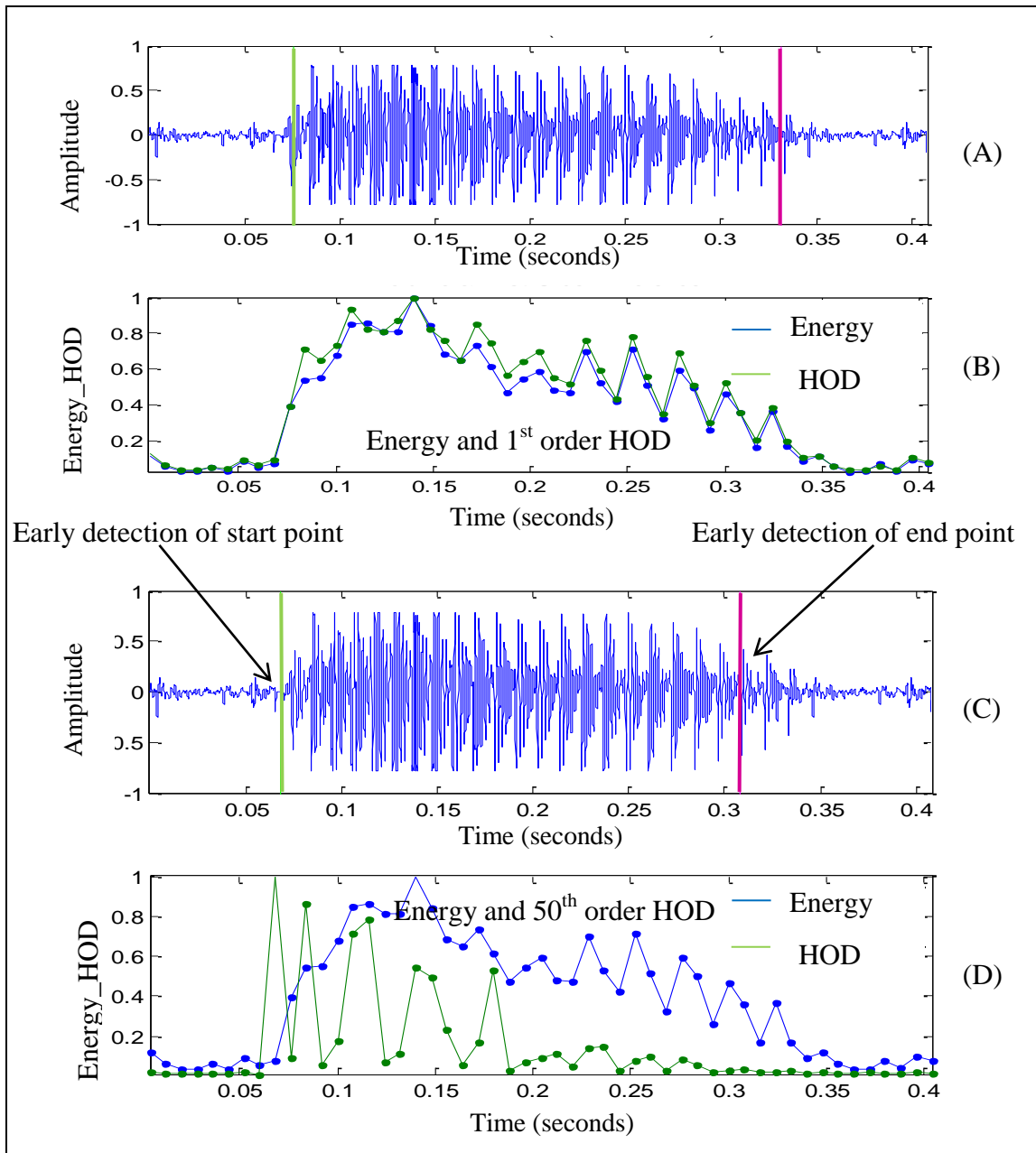


Figure 3.16: Effect of different order of HOD algorithm on VAD

VAD is performed on the pre-processed filled pause and elongation datasets (i.e FP_DATA and ELO_DATA) using energy, ZCR and HOD methods. Using combination of these features, six filled pause and elongation datasets are produced as in Table 3.6. The name of each dataset refers to the chosen VAD method. VAD_01, VAD_02 and VAD_03 are specifically for filled pause datasets, while VAD_04, VAD_05 and VAD_06 are designated for elongations datasets.

Table 3.6:
Disfluencies datasets for filled pause and elongation

Datasets	Descriptions of the datasets
VAD_01	Automated VAD of filled pause using energy
VAD_02	Automated VAD of filled pause using energy + ZCR
VAD_03	Automated VAD of filled pause using energy + Higher order differences
VAD_04	Automated VAD of elongation using energy
VAD_05	Automated VAD of elongation using energy + ZCR
VAD_06	Automated VAD of elongation using energy + Higher order differences

The reduction percentage is calculated for each VAD method applied in this research. The reduction percentage is calculated as in equation (3.19).

$$\frac{Vb - Va}{Vb} \times 100 \quad (3.19)$$

where,

Vb = Value of FF1 before energy and ZCR-based VAD is applied

Va = Value of FF1 after energy and ZCR-based VAD is applied

Results and discussion of VAD is postponed to Chapter 6. Based on VAD experiments, a new filled pause and elongation datasets are produced. These datasets are further used for feature extraction and pattern classification.

3.6 SUMMARY

This chapter presents the dataset creation process for filled pauses (i.e. FP_DATA) and (i.e. ELO_DATA) elongation of Malay language disfluencies which denotes the first contribution of this thesis. To create the data collection, the need to understand Malay language speech sounds and rules is required since the spoken filled pause and elongation is language dependent. A closer look at Malay language filled pause and elongations are also done. A summary of their characteristics are listed as follows:

- i. Duration of all filled pause and elongations are greater than 200ms.
- ii. Filled pauses have flat fundamental frequency and stable formant frequency.
- iii. All filled pauses consist of phonemes which is pronounced unvaried, while all elongations comprise consonant and vowels causing expressive intonation.
- iv. Unlike English and Mandarin language, all elongation occurred at the last syllable of normal words.

The characteristics of filled pauses and elongations identified in this chapter are important to further recommend the suitable acoustical feature representation. Chapter Four continues with the discussion of feature extraction and the proposed feature extraction technique employed in this thesis for classification of filled pause and elongation.

This chapter also discussed the overview of the research methodology as well as the pre-processing that need to be taken on the datasets. The manually segmented FP_DATA and ELO_DATA are pre-processed using standard methods of normalization, pre-emphasis, framing and windowing. Voice activity detection (VAD) follows to get the accurate representation of filled pause (i.e. FP_DATA) and elongations (i.e. ELO_DATA). These datasets are further used in acoustical feature extraction conferred in Chapter Four of this thesis. Chapter Five and Six conclude the overall methodology by elaborating classification and evaluation of filled pause and elongation.

CHAPTER FOUR

FEATURE EXTRACTION FOR NEW ACOUSTICAL FEATURE CONSTRUCTION

4.1 INTRODUCTION

This chapter examines acoustical feature extraction to obtain a feature representation for the Malay language filled pause and elongation. As stated in the problem statement, the use of existing speech energy extraction technique is unable to represent filled pause and elongation discreetly. Therefore, the main aim of this chapter is to produce a new method of extracting speech energy to get a more accurate energy representation of filled pause and elongation (i.e. objective two). Overall, the process of acoustical feature extraction is illustrated in Figure 4.1. This process aims to collect a compilation of normalised acoustical feature vectors that are extracted from each speech segment (FP and ELO) to be used later in the classification stage.

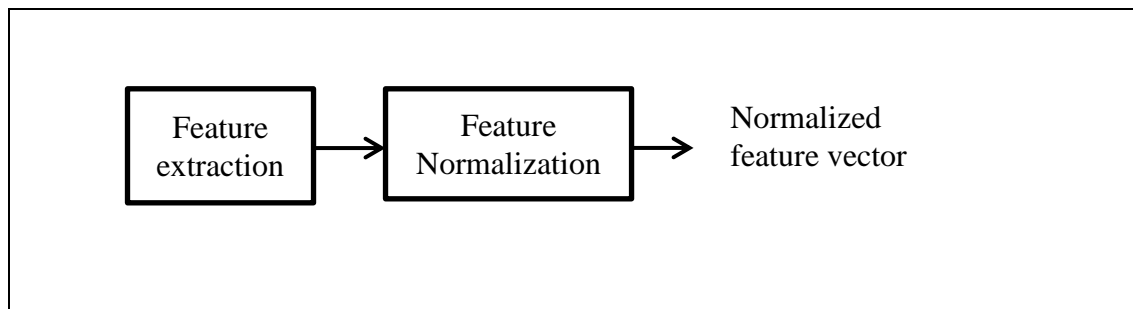


Figure 4.1: Acoustical feature extraction process for each speech segment

4.2 STANDARD FILLED PAUSE ACOUSTICAL FEATURES

Previously, several well-established acoustical features were used for research regarding filled pauses in speech events. The well-established acoustical features of filled pause and elongation (Audhkhasi & Angeles, 2009; Veiga, 2011; Verkhodanova & Shapranov, 2014) were extracted in this research in order to see the effectiveness in filled pause and elongation classification. The features consist of Fundamental Frequency (F0), Short Time Energy (STE), Formant Frequency (FF), Mel Frequency Cepstral Coefficients (MFCC) and Zero Crossing Rates (ZCR).

4.2.1 Formant Frequency

Feature extraction starts with one of the most common acoustical features of filled pause and elongation, formant frequency. In this research, the formant frequencies were obtained from the poles of a linear prediction model of speech by using linear prediction coding (LPC). The aim was to estimate the transfer function of the vocal tract from the speech from the source signal as modeled in equation (4.1).

$$s(n) = - \sum_{i=1}^p a_k \cdot s(n-k) + e(n) \quad (4.1)$$

where,

p = number of coefficient in the model

a_k = p^{th} order of the linear prediction coefficients

In the z -transform domain, a linear prediction model of speech $X(z)$ is expressed as in equation (4.2).

$$X(z) = E(z)V(z) \quad (4.2)$$

where,

$X(z)$ = speech signal

$E(z)$ = z -transform the speech signal

$V(z)$ = z -transform of the linear prediction model

The LP model $V(z)$ can be expressed as a cascade combination of a set of second order resonators and a first order model as in equation (4.3).

$$V(z, m) = G(m) \frac{1}{1 + r_0(m)z^{-1}} \prod_{k=1}^{P/2} \frac{1}{1 - 2r_k(m) \cos(\varphi_k(m))z^{-1} + r_k^2(m)z^{-2}} \quad (4.3)$$

where

$r_k(m)$ = time-varying radii of the LP model poles

$\varphi_k(m)$ = angular frequencies of the LP model poles

$P+1$ = LP model order

$G(m)$ = gain of the LP model for frame m .

Next, the poles that were obtained from the LP model were considered to be formant candidates (Kim et al., 2006; Yan et al., 2007). The formant frequency and bandwidth of the formant are represented as in equation (4.4) and (4.5).

$$F_k = \frac{fs}{2\pi} \varphi_0 \quad (4.4)$$

$$B = -\frac{fs}{\pi} \ln(r_0) \quad (4.5)$$

where,

- r_0 = Magnitude of the LP model pole
- fs = Frequency sampling
- F_k = Formant frequency
- B_k = 3-db formant bandwidth

Each extracted formant frequency consists of rows of vectors as in equation (4.6).

$$FF_y = [v_1 \ v_2 \ v_3 \ v_4 \ v_5 \dots \ v_l] \quad (4.6)$$

The length of the vectors is proportional to the number of frames in the speech. Each formant frequency vectors is represented in a form of standard deviation (std_dev) to get only one vector from it as shown in equation (4.7).

$$FF_y = \sqrt{\frac{\sum (v - v)^2}{n - 1}} \quad (4.7)$$

Where FF = Formant frequency of a filled pause or elongation

y = The level of formant frequency that is being measured

v = the formant frequency vectors

As an example in Figure 4.2 shows, the standard deviation of FF1 of the elongated word /DI/ is calculated from 24 vectors and becomes only one scalar (27.47) when calculated using equation (4.7). This post-processing for formant frequency representation is a standard process as was done in other filled pause analysis research (Audhkhasi & Angeles, 2009; Kaushik et al., 2010; Veiga, 2011). The standard

deviation is calculated as a measurement of formant frequency stability. Standard deviation is also used in Temko et al., (2008) to get one representation of several acoustical features, inclusive of zero crossing rates, fundamental frequency, and energy.

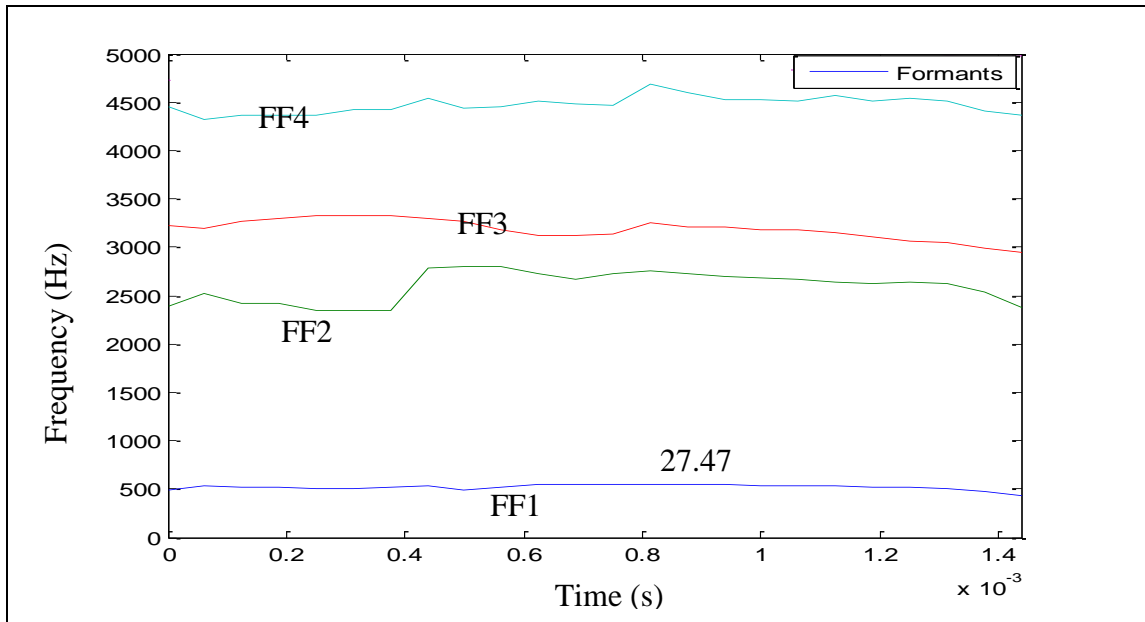


Figure 4.2: Example of formant frequency plot of elongation of the word /DI/

Post-processing is also done for the other formant frequencies (FF2, FF3, and FF4). Previously, formant frequency was used by (Audhkhasi & Angeles, 2009) and (Kaushik et al., 2010) in filled pause research. This feature has shown its performance in representing filled pause. However, in their research both disfluencies are grouped into the same class due to the similar acoustical features' pattern of filled pause and elongation. Compared to this research, both filled pause and elongation are treated distinctly. Due to the popularity of formant frequency in filled pause research, this feature is taken into account in this research in order to test its performance.

4.2.2 Fundamental Frequency

Fundamental frequency (F0) is a popular acoustical feature used in previous filled pause research. Theoretically, the minimal intonation during filled pause and elongation causes F0 to remain almost flat (Audhkhasi & Angeles, 2009). F0 is always correlated with pitch in many pitch measurements, because F0 can give a

measurement of pitch virtually or graphically. Pitch is an auditory percept of human tone (Talkin, 1995). In this research, the common fundamental frequency detection method utilising auto-correlation function (ACF) is used. The aim of ACF is to find the peak value of the autocorrelation in the region of interest. For a discrete time signal $y(n)$, the long time ACF function is defined as shown in equation (4.8).

$$\phi_y(m) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N y(n)y(n+m) \quad (4.8)$$

Basically, the ACF of a signal is a non-invertible transformation of the signal. It is useful for presenting the structure in the signal waveform. By assuming $y(n)$ is exactly periodic with period P , i.e. the ACF is also periodic with the same period as in equation (4.9).

$$\phi_y(m) = \phi_y(m+P) \quad (4.9)$$

The periodicity of the ACF shows the periodicity of the speech signal. However, a speech waveform is non-stationary, thus a long-time ACF as in (4.9) will not give an accurate F0 estimation. The convenient way to address this is to define the short-time ACF to enable F0 estimation on every short segments of the speech. The short-time ACF is defined as in equation (4.10).

$$\phi_y(m) = \frac{1}{N} \sum_{n=0}^{N-1} [y(n+\ell)w(n)][y(n+\ell+m)w(n+m)] \quad (4.10)$$

$$0 \leq m \leq M_0 - 1$$

where

$w(n)$ = Hamming window

ℓ = The starting sample of the frame index

N = The analyzed section length

N' = Number of signal samples used in the computation of $\phi_y(m)$

M_0 = number of computed autocorrelation points

The process of F0 detection is illustrated as in Figure 4.3.

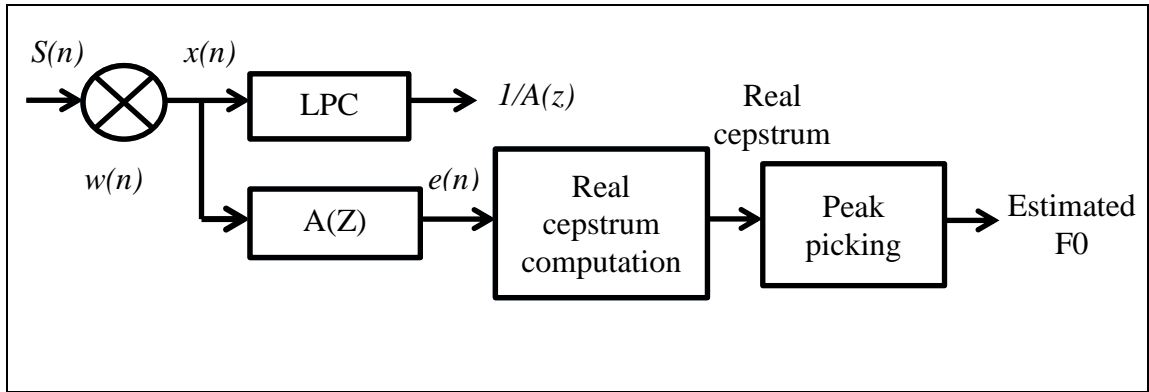


Figure 4.3: Process of F0 detection by using autocorrelation function

From Figure 4.3, the F0 detection processes are discussed according to the following steps:

- i. The source signal $s(n)$ first went through pre-processing to be chopped into blocks of a short time signal $x(n)$ by using framing and a windowing process, as in *Section 3.5*.
- ii. The coefficients $\frac{1}{A(z)}$ of all-pole vocal tract model for each $x(n)$ are estimated by using the LPC method (*refer to Section 4.2.1*).
- iii. The inverse transfer function of $\frac{1}{A(z)}$ which is $A(z)$ results in residual signal $e(n)$.
- iv. The real cepstrum of the resulting residual signal $e(n)$ is then calculated.
- v. Finally, the peaks of the real cepstrum denotes the fundamental frequency (Garg et al., 2011).

The same procedure is taken as that of formant frequency. The standard deviation is calculated on each speech sample, F0. The process of standard deviation computation is implemented to get a single representation of the F0.

4.2.3 Mel-Frequency Cepstral Coefficients

The next acoustical feature that was extracted in this research was Mel frequency cepstral coefficients (MFCC). It was chosen in this research due to its equally spaced frequency based on Mel-scale, which is close to the human auditory system response (Holmes & Holmes, 2001). It is also related to the speech energy. In filled pause research, MFCC is a preferred feature used in analysis (Li et al., 2008). MFCC is a dominant data representation in speech recognition and synthesis, especially for a short-term power spectrum. The process of MFCC extraction is illustrated as in Figure 4.4.

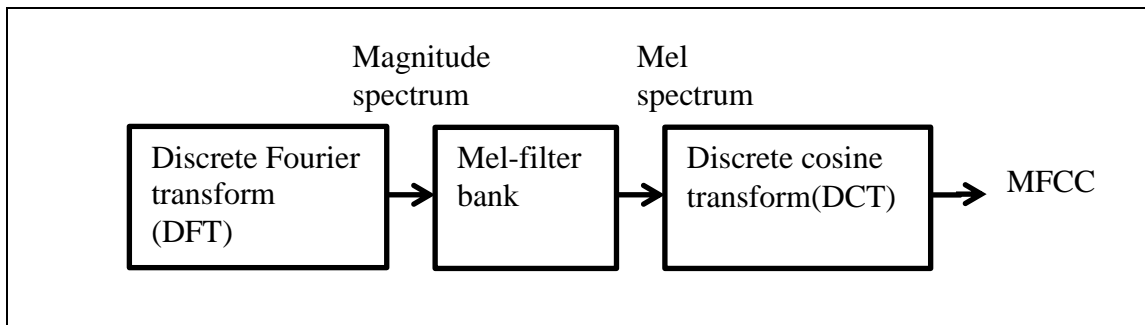


Figure 4.4: Mel-frequency cepstral coefficients extraction process

There are two stages involved in MFCC extraction. The first stage is to calculate the cepstrum and the second stage is ‘Mel’ scaling. By using ‘Mel’ as the unit reference, 1000Hz is written as 1000 Mel. The steps were as follows:

Step 1: Calculation of overall speech energy

The energy of a speech signal or a frame of the speech signal $s(n)$ is calculated as equation (4.11).

$$S_i = |S_k|, \text{ for } i=0,1, \dots, N/2 \quad (4.11)$$

where,

$S(k)$ is the N -point Discrete Fourier Transform (DFT) of the speech signal or a frame of the speech signal as equation (4.12). DFT is applied to convert each time domain frame of N samples into a frequency domain. DFT is computed via fast Fourier Transform (FFT).

$$S(k) = \sum_{n=0}^{N-1} s(n)e^{j2\pi kn/N}, \text{ for } k=0,1,\dots,N-1 \quad (4.12)$$

Step 2: Calculation of critical band energy

Next, the energy in each critical band is obtained by applying the conceptual triangular windows to the spectral magnitude in equation (4.13).

$$E_j = \sum_{i=0}^{(N/2)-1} |S_i \cdot h_j(i)|, \text{ for } i=0,1,\dots,J \quad (4.13)$$

where,

J = total of triangular filters, $h_j(i)$

Step 3: Conversion of linear scale frequency to Mel scale

The scale of the frequency is converted from linear to Mel scale since the results of DFT (in step 4.2.3.1) produced a wide frequency range. Theoretically, human speech does not follow the linear scale. The 'Mel' for a given frequency is calculated by using equation (4.14).

$$Cs(n) = \sum_{i=1}^J \log_{10}(E_j) \cos\left[n\left(j + 0.5\right)\frac{\pi}{j}\right] \quad (4.14)$$

where,

n = number of MFCC coefficients to be gathered. Normally the coefficients are set from 8 to 14 (Deng & O'Shaughnessy, 2003).

Step 4: Application of logarithm on each Mel scale

The logarithm is applied using 13-cepstral coefficient of MFCC. The first coefficient $cs(0)$ represents the average power in the speech signal. However, $cs(0)$ is not often used in recognition applications since the average power varies considerably depending on the recording channel. The coefficients $Cs(n)$ give increasingly finer

spectral details for each $n > 1$ (Deng & O'Shaughnessy, 2003). Therefore, 13-MFCC is chosen because as the order of index number increases, the spectral details become negligibly small.

Step 5: Conversion from log Mel spectrum to time domain

The log Mel spectrum is converted back to the time domain by using Discrete Cosine Transform (DCT).

4.2.4 Zero Crossing Rates

Short time zero crossing rates (ZCR) is preferred due to its ability in providing a different distribution for voiced, unvoiced, silent, consonant and vowel parts of the speech. The filled pause is in the form of vowels, while elongation is the combination of a vowel and consonant. Theoretically, ZCR for a vowel is lower due to voiced characteristics of vowel segments, while the ZCR for elongation is higher due to the existence of a consonant at the beginning of the word. Therefore, the ZCR differences between filled pause and elongation due to the consonant and vowel characterisation are calculated. The ZCR is calculated as in equation (3.5) and the detailed process of extraction is discussed in Chapter Three (*refer Section 3.5.5.2*). The difference of ZCR utilisation between pre-processing (Chapter Three) and this section is that the ZCR is used here as the feature to discriminate between filled pause and elongation. Alternatively, in Chapter Three, it is used as a threshold method in detecting the starting and endpoint of the speech for voiced activity analysis. The output of ZCR calculation from equation (3.15) is the rates of zero crossing for each speech sample segment of filled pause and elongation. For this section, the ZCR is taken as the average value of the ZCR rate of each frame.

4.2.5 Energy

The next extracted acoustical feature is energy. Energy is widely used in filled pause research (Gaurav & Nigel, 2006; Li et al., 2008; Veiga, 2011). The use of energy can be found in different language of filled pause studies such as Mandarin,

European Portuguese and English. Since filled pause and elongation is language specific (Zgank & Maucec, 2010), the performance of energy were reported differently. It was proven in Veiga et al., (2011) that energy is unable to differentiate filled pause and elongation of European Portuguese language due to the equal pattern of energy stability. In contrast with Li et al., (2008), the energy along with MFCC and F0 have shown promising classification performance for Mandarin filled pause and elongation classification.

In Chapter Three, energy is used as a threshold method for voice activity detection. On the other hand, this chapter used energy is used as one of the acoustical features to be classified as either filled pause or elongation. In general, the process of getting the representation of each speech sample's energy is by using the standard method (Jalil et al., 2013) that is by calculating the sum of the energy of each short speech frame as in equation (3.8) (*Section 3.5.5.1*).

The next step is to calculate the energy's standard deviation of the whole speech segment to measure the energy's stability. Energy standard deviation of the filled pause is expected to be small (Veiga, 2011) compared to other utterances in a spontaneous speech sentence as they are presumed to be more stable. Energy example of filled pause and elongation is taken to demonstrate its function in representing elongation and filled pause as shown in Figure 4.5 and Figure 4.6.

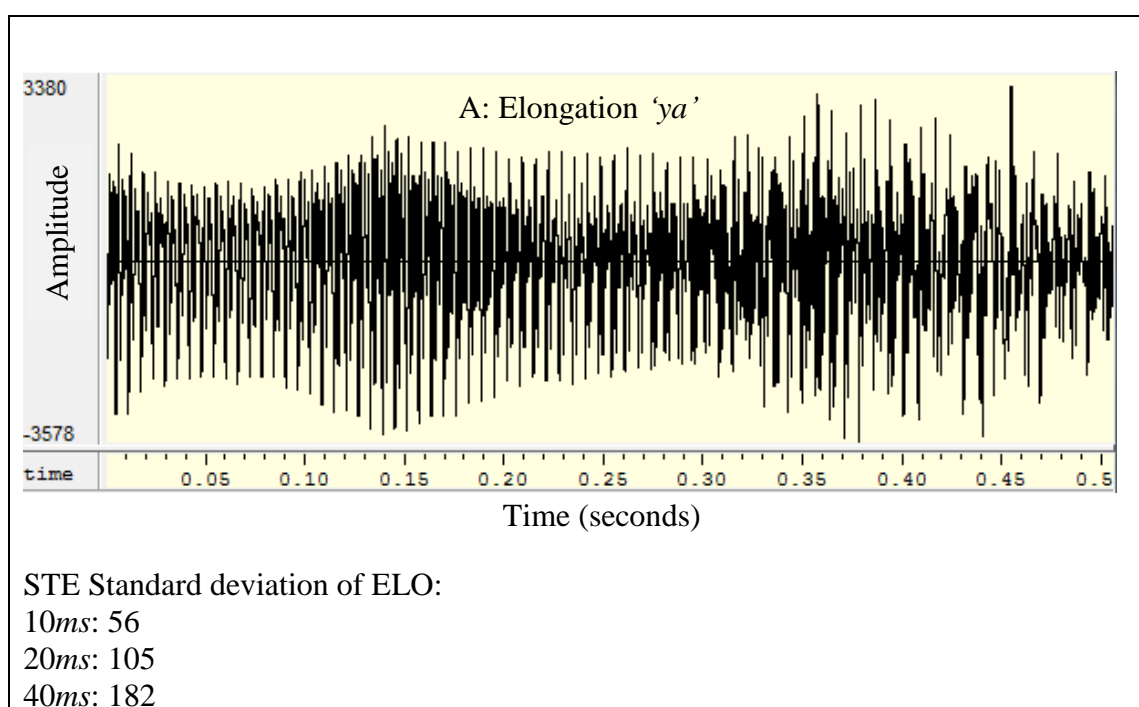


Figure 4.5: Example of STE measurements on elongation

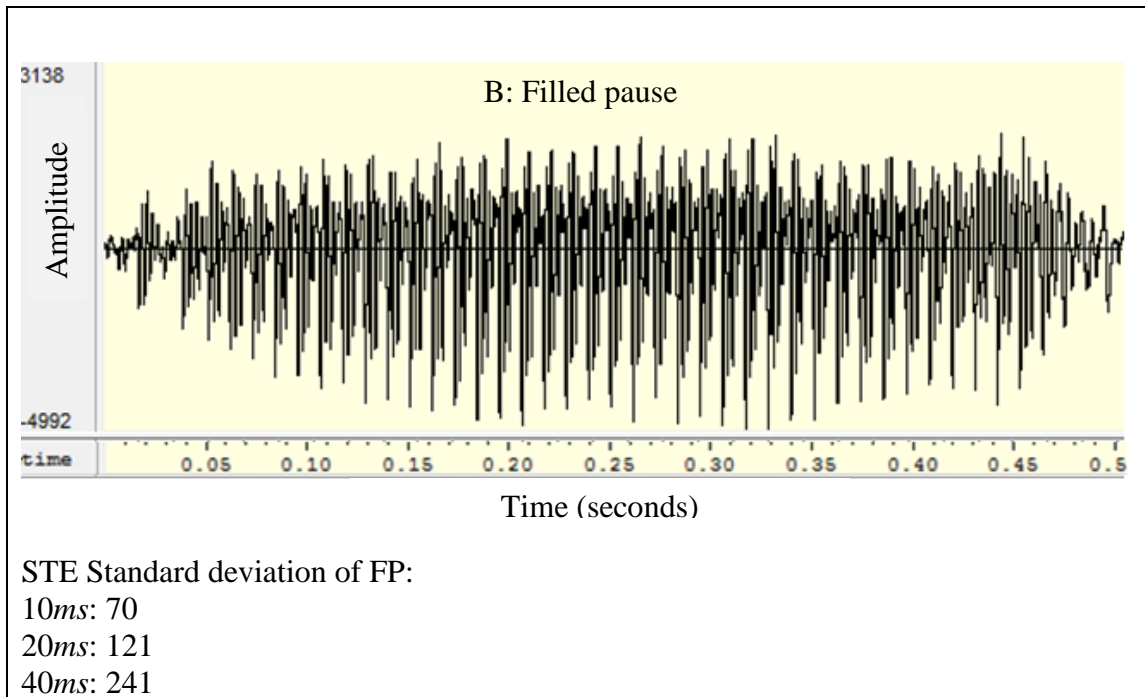


Figure 4.6: Example of STE measurements on filled pause

Hypothetically, elongation (A) in Figure 4.5 should produce higher STE, while filled pause (B) in Figure 4.6 lower due to the stability assumption of filled pause. However, both filled pause and elongation produced the contradicting value of STE. For each duration that are tested (i.e. 10ms, 20ms and 40ms), the standard deviation of the energy produced by elongation are denoted as 56, 105 and 182 which are lower compared to filled pause energy's standard deviations (i.e. 70, 121, 241).

The distribution of energy value of both filled pause and elongation is shown in Figure 4.7. From Figure 4.7, it is obviously seen that the energy representation (energy standard deviation) of filled pause and elongation is overlapping. It shows that the filled pause and elongation cannot be differentiated by using energy as the feature. In filled pause research, energy is an important feature. Several acoustical features that were previously tested in filled pause classification such as fundamental frequency and spectral envelope are correlated with energy (Rosenberg & Hirschberg, 2006). Generally, the energy of filled pause is stable and constant, as proven in Gaurav & Nigel, (2006). However, due to the transition between consonant and vowel in the elongation, the standard method of energy measurement is not able to represent this transition named as expressive intonation. Therefore, another way of exploiting the energy of the speech is by using the local information of the speech energy need to be investigated. This is further explored and discussed in the next subsection.

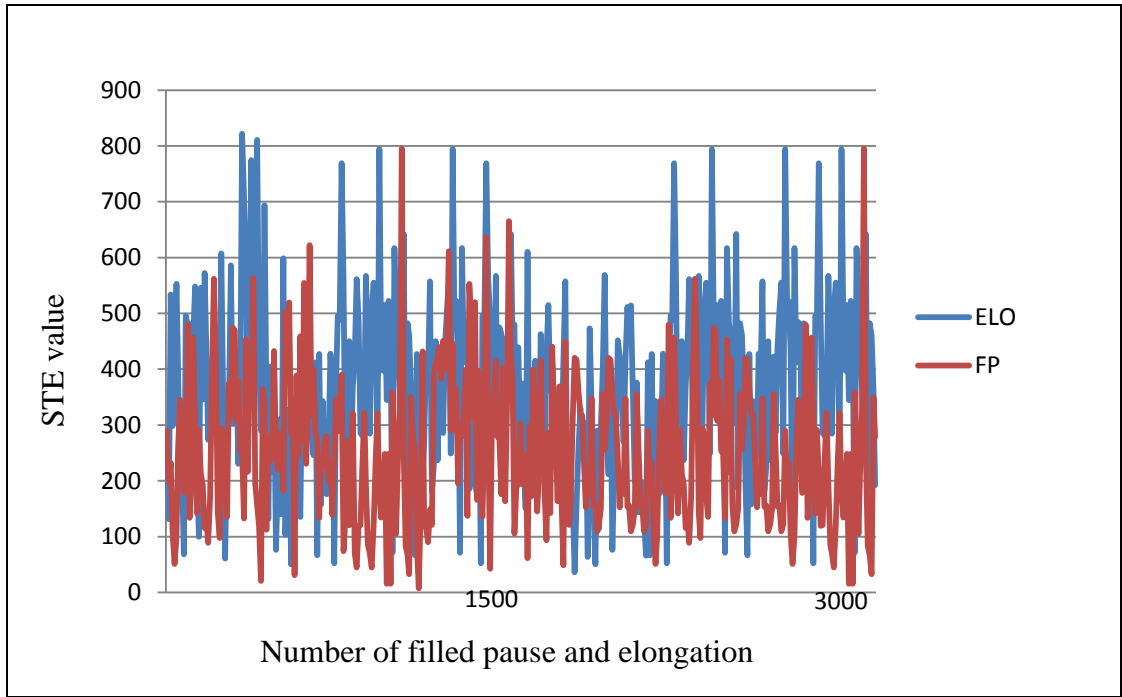


Figure 4.7: STE value distribution for filled pause and elongation

4.3 LOCAL MAXIMA OF THE SPEECH ENERGY

This research proposed another technique to measure the energy of the disfluencies by exploiting the local information of the speech signal. Local maxima are the highest points marked at an interval as shown in Figure 4.8. It is observable that the local maximum and minimum is situated in the upper and lower plane respectively. The local maxima are also the combination of the local maximum from one to another interval of a function or signal. For a point to be declared as the local maximum, of a function $F(x)$, it must be greater than or (equal to) the height anywhere else in that interval as describe in equation (4.15).

$$P_{LM} = P \geq hs \quad (4.15)$$

where,

P_{LM} = Point extracted as local maxima

P = any point in a speech

hs = any height in a speech

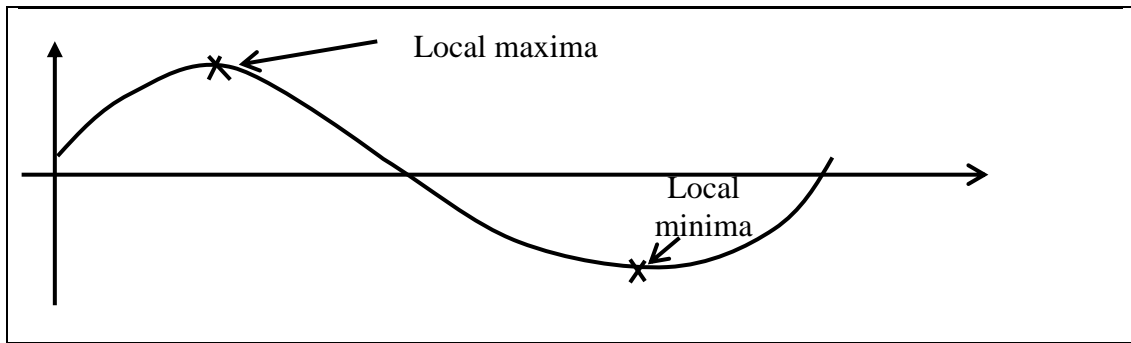


Figure 4.8: Local maxima location

It is well-known that the energy of filled pause is stable and constant due to its stable pattern of unvaried phoneme pronunciation (Veiga, 2011). Previously, the speech energy is calculated by firstly framing it into small portions and the energy is calculated on the basis of total energy of each portion. Then, the standard deviation is calculated to measure the stability of the energy. However, as shown in Figure 4.7, the range of overlapping STE values cannot significantly separate filled pause and elongation as two classes. The transition between consonant and vowel in the elongation is minute; it requires another method of representation to measure the local information.

Speech energy is closely related to the amplitude of the speech (Izzad et al., 2013). Instead of calculating the total energy of each frame, the energy stability of the speech in the proposed technique is measured based on the amplitude transition from one frame to another. To measure the amplitude transition, this research proposed the exploitation of the local maxima points of the speech. Previously, several techniques of local maxima extraction have been proposed for various domain. Basically, the techniques of local maxima extraction depend on the threshold parameter selection. One of the techniques of local maxima extraction is by utilizing the distance between peaks as threshold (Cheng et al., 2015). The other technique is performed by using minimum height (Bertot et al., 2014) as threshold. Both techniques are discussed in the following subsections.

4.3.1 Local Maxima Extraction Using Minimum Peak Distance Threshold

The technique of local maxima extraction by using minimum peak distance was used in Cheng et al., (2015). The steps of local maxima by using minimum peak distance are as follows:

Step 1: Specify a positive value for a minimum peak distance as a threshold.

This threshold is referred as minimum frame.

Step 2: Identify all the peaks within the specified minimum frame.

Step 3: Arrange the peaks of the data in descending order.

Step 4: Choose the highest peak among the identified peaks in the first frame of the specified minimum distance as the first local maxima.

Step 5: Repeat steps (2 to 3) for the next consecutive frames of minimum distance to extract the next local maxima until there is no peak to be considered.

This technique is applied in this research. However, the limitation of this technique is its inability to detect certain points that should be declared as local maxima points. In the example of Figure 4.9, a random threshold of '2' is applied as the minimum peak distance. The expected result of local maxima is '4' and '7'. However, local maxima by using minimum peak distance only displayed '7' as the local maxima.

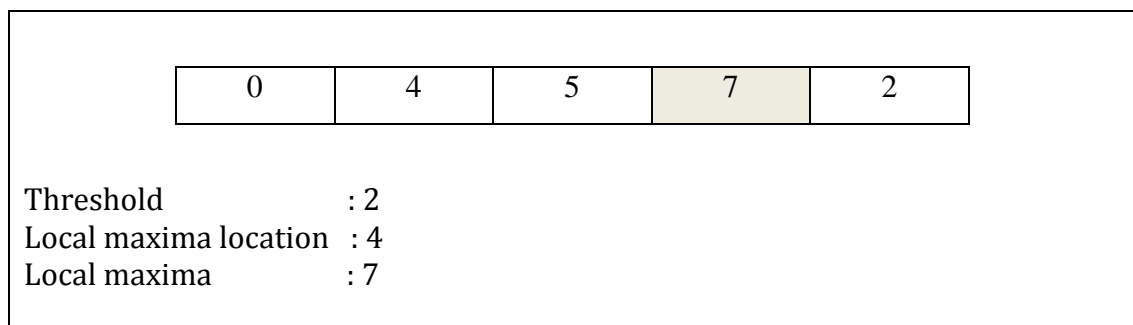


Figure 4.9: Example of local maxima extraction by using minimum peak distance as threshold

The distribution of the local maxima extraction by using minimum peak distance is shown in Figure 4.10. It indicates that local maxima extraction using the local maxima values of filled pause and elongation are overlapping. It can be seen that, local maxima extraction using minimum peak distance is unable to segregate filled pause and elongation.

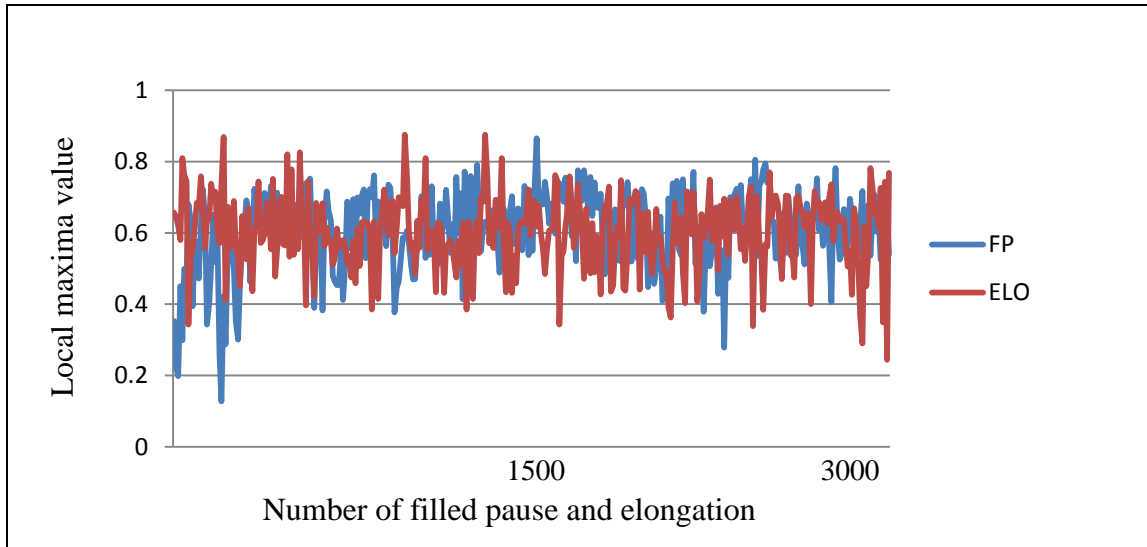


Figure 4.10: Distribution of local maxima using minimum peak distance

4.3.2 Local Maxima Extraction Using Minimum Peak Height Threshold

Bertot et al., (2014) used the height of the speech signal as a threshold. In this technique, the peak is detected by first order difference information. A peak occurs when the trend changes from upward to downward, i.e., a peak is where the difference changed from a streak of positives and zeros to negative. The steps involved in this local maxima extraction using minimum peak height threshold are as follows:

Step 1: Specify a positive value for a minimum peak height as a threshold, *thre*. The threshold is defined as in equation (4.16).

$$thre = \alpha \frac{1}{Ne} \sum_{i=0}^{Ne-1} e_i^2 \quad (4.16)$$

where,

Ne = the number of sample

e = the signal's sample

α = constant for threshold adjustment

Step 2: Compare each peak with the *thre*

Step 3: Choose peak that exceeded the *thre* as the first local maxima

Step 4: Repeat step (2 to 3) for the next consecutive point of minimum height to extract the next local maxima until there is no peak to be considered.

When the same threshold (i.e Figure 4.9) of ‘2’ is used, the same local maximum as the previous method is selected that is ‘4’. The distribution of the local maxima extraction by using minimum peak height is shown in Figure 4.11. It is obvious that the local maxima value between filled pause and elongation are overlapped. It shows that local maxima extraction by using minimum peak height is not suitable to classify filled pause and elongation.

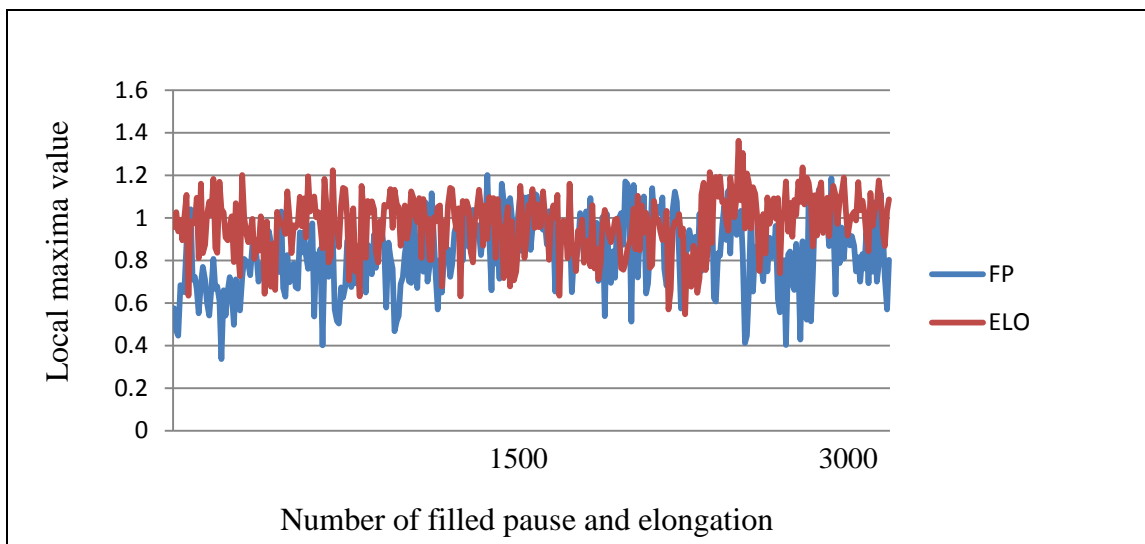


Figure 4.11: Distribution of local maxima using minimum peak height

4.3.3 Proposed Local Maxima Threshold Selection

To overcome the limitation of the previous local maxima techniques, this research proposed an improved local maxima of the speech energy technique referred to as (LM-E), hereafter. The LM-E is executed by directly comparing one peak point to another using different threshold selection. In this proposed method, different adjustable positive scalar number is tested as threshold to observe the most suitable parameter. By using the same example in Figure 4.9, the proposed algorithm successfully extracted the expected local maxima as shown in Figure 4.12.

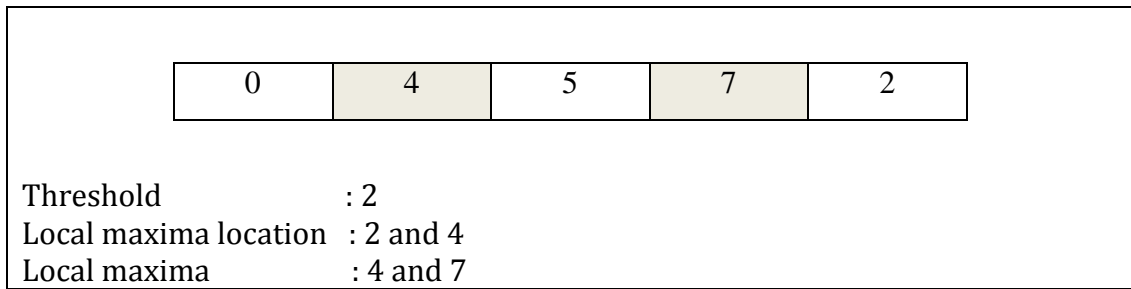


Figure 4.12: Example of proposed local maxima extraction

The LM-E is executed by setting a threshold as height difference. The steps are as follows:

Step 1: Specify a positive value for a height difference between peaks as a threshold *thr*.

Step 2: Compare each peak with its neighbouring peaks.

Step 3: If the height difference between peaks is higher than the *thre*, returns the peak as local maxima.

Step 4: Repeat steps (2 to 3) for the next consecutive point of minimum height difference to extract the next local maxima until there is no peak to be considered.

A randomly chosen filled pause (FP01.wav) and elongation (ELO01.wav) are taken as examples as shown in Figure 4.13 to test the proposed LM-E effectiveness. The standard deviation of all three LM-E methods using default threshold for FP01.wav and ELO01.wav are calculated for a fair comparison. Results are presented in Table 4.1. As stated previously, the constant and stable energy of filled pause should produce lower standard deviation value compared to elongations. From the results shown in Table 4.1, the LM-E using minimum peak distance and minimum peak height threshold produces higher standard deviation of filled pause as compared to elongation. However, the proposed LM-E threshold selection adheres to the assumption that standard deviation of filled pause (i.e. 0.553) is lower than elongation (i.e. 0.695).

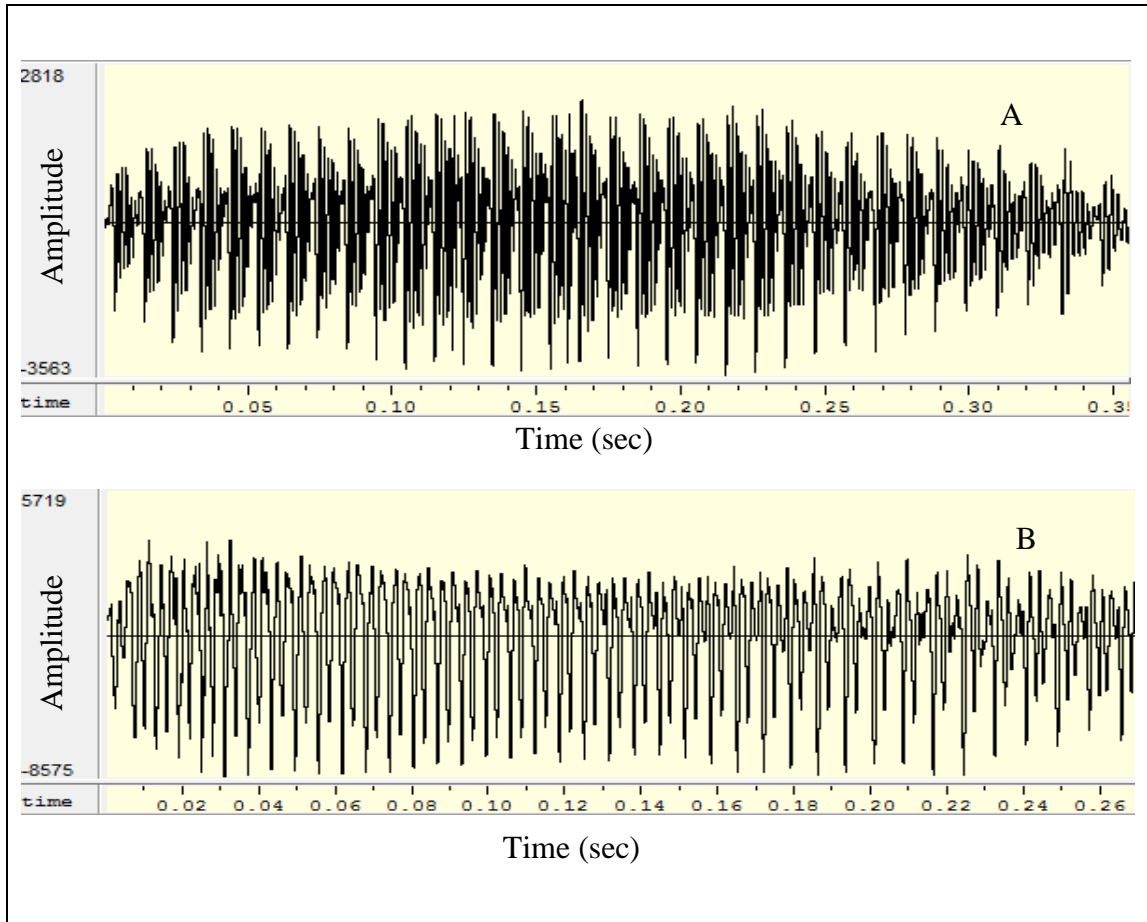


Figure 4.13: Filled pause (A) and elongation (B) from 28082008 dataset

Table 4.1:

Standard deviation comparison between LM-E techniques

Techniques	Threshold parameter	Filled pause	Elongation
Minimum peak distance as threshold	Default: 1	0.846	0.761
Minimum peak height as threshold	Default: -Inf	0.846	0.761
Proposed LM-E	Default: 0.001	0.553	0.695

The distribution of value for the proposed LM-E is shown in Figure 4.14. Compared to STE (i.e. Figure 4.7), LM-E using minimum peak distance (i.e. Figure 4.10) and LM-E using minimum peak height thresholds (i.e. Figure 4.11), the proposed LM-E (i.e. Figure 4.14) showed better discrimination of both filled pause and elongation.

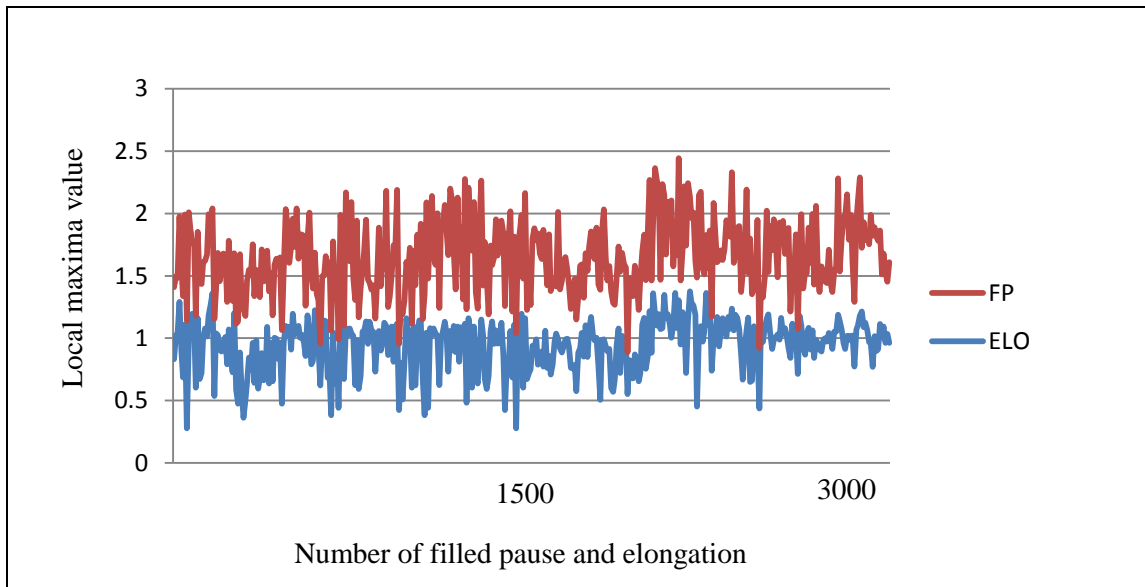


Figure 4.14: Proposed LM-E data distribution of filled pause and elongation.

Even though the proposed LM-E technique showed promising, potential in discriminating filled pause and elongation, further investigation on the thresholding technique is done. The aim of the investigation is to find a suitable means of calculating the threshold for LM-E's extraction. Two thresholds methods are applied as follows:

i) Default value of 0.001

Based on observations of the extracted local maxima points that are extracted, the more local maxima points are extracted when lower threshold is used while less local maxima points are extracted when higher threshold is extracted.

ii) Mid-range of the disfluencies

The statistical measurement is used to get the representation of the speech amplitude that is related to the threshold. This research tested the average, midrange and standard deviation of both filled pauses and elongations. The midrange produces better classification accuracy as compared to the other statistical measurement.

To test the thresholds methods of default value and midrange, a classification process is executed. The classification process is discussed in Chapter Five. However, when default value of 0.001 is applied, the classifier's accuracy produced better result only for elongation. On the other hand, the accuracy becomes higher for filled pause when the midrange is used. Therefore, the advantage of default value and midrange

are applied to produce better performance for both classes of disfluencies. The midrange value of filled pause and elongation is calculated as in equation (4.17).

$$midR = \frac{\max(e) + \min(e)}{2} \quad (4.17)$$

where

e = the speech's sample

iii) Rule-based threshold

Based on the threshold investigations, a rule-based threshold is proposed to calculate the LM-E. The average midrange value of all filled pauses and elongations are 0.4 and 0.2, respectively. It can be seen that the average midrange of the elongation is lower as compared to the filled pause. Therefore, the research chose 0.4 as the midrange benchmark. The rules tested on the algorithm are as follows:

a) Rules_01

If $midR \geq 0.4$
 Detect LM-E using midrange as threshold;
 Else
 Detect LM-E using default value as threshold;

b) Rules_02

If $midR < 0.4$
 Detect LM-E using midrange as threshold;
 Else $m > 0.4$
 Detect LM-E using default value as threshold;

The overall flow chart of the proposed local maxima threshold selection method is shown in Figure 4.15.

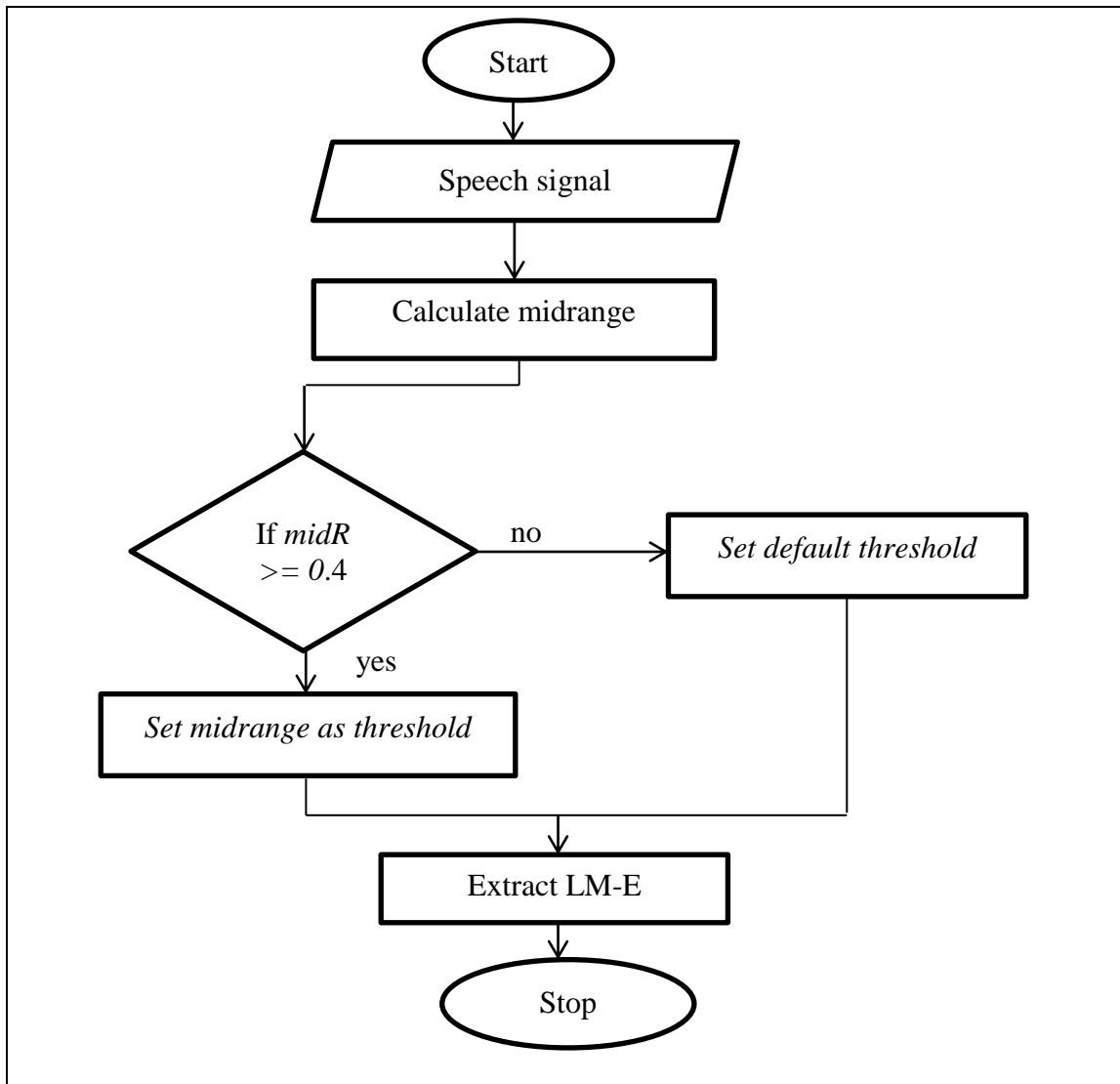


Figure 4.15: Flowchart of the proposed LM-E technique

The same filled pause and elongation as in Figure 4.5 and Figure 4.6 are taken to be compared with the proposed LM-E. It can be seen that the proposed LM-E produced lower LM-E standard deviation for filled pause (i.e. 0.63) and compared to elongation (i.e. 0.97) as shown in Figure 4.16. As stated earlier, the standard energy calculation of the filled pause is done by calculating the sum of the energy of each short speech frame. When compared with the standard energy calculation, this proposed LM-E gives correct representation of filled pause and elongation (i.e. lower standard deviation of LM-E for filled pause and higher LM-E for elongation).

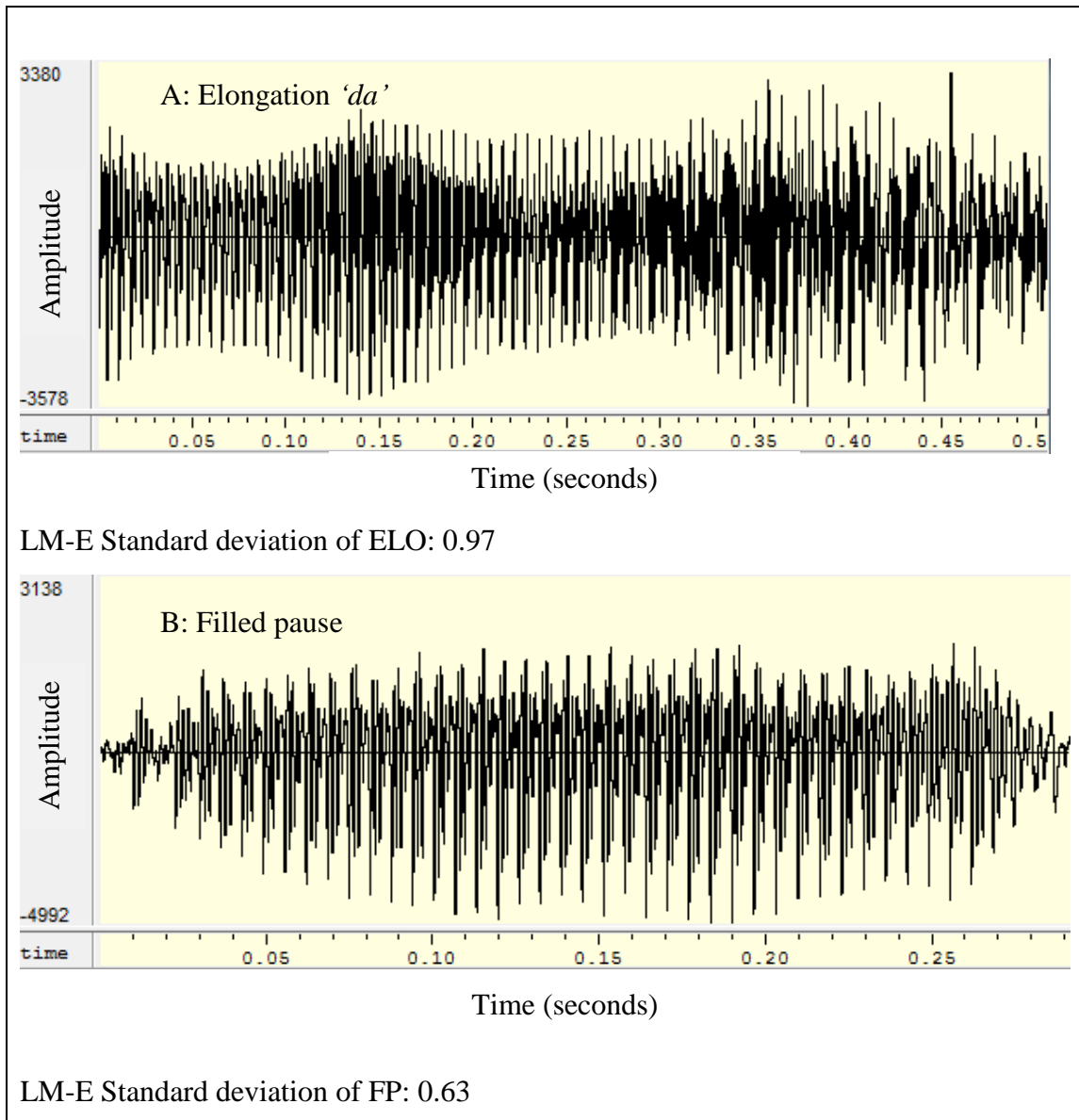


Figure 4.16: Example of LM-E measurements on filled pause and elongation

4.4 FEATURE RANKING

Feature ranking is executed to get the prior knowledge of feature importance rank in representing the filled pause and elongation. The ranking is also done to help identify which acoustical features are most suitable for Malay language filled pause and elongation. All the acoustical features described earlier in this chapter are well-established, in field of filled pause study. However, they have never been simultaneously analysed to observe its importance in classifying filled pause and elongation. Overall, nine acoustical features are extracted in this research, and each feature comprises feature values that can be categorised into two classes i.e. filled

pause and elongation. Each acoustical feature that was extracted has its own notation as presented in Table 4.2. This notation is used in this thesis to simplify their presentation.

Table 4.2:
The acoustical features and its notations

No	Acoustical Feature	Notation
1	Formant frequency1	FF1
2	Formant frequency2	FF2
3	Formant frequency3	FF3
4	Formant frequency4	FF4
5	Fundamental frequency	F0
6	Short time energy	STE
7	Mel frequency cepstral coefficients	MFCC
8	Local maxima of the speech energy	LM-E
9	Zero crossing rates	ZCR

A Random Forest based algorithm called Boruta algorithm is chosen to calculate and define the most important attributes based on their importance measure. Boruta was developed by Kursa & Rudnicki, (2010). The features are ranked from the most important to the least by using Z-score. Boruta algorithm has received a lot of attention in numerous areas and the stability is proven Kursa, (2014). Previously, in Kursa, (2014) the Boruta algorithm has been used and compared with the existing Random-Forest algorithm such as Artificial Contrasts with Ensembles (RF-ACE), Recursive Feature Elimination (RFE) and Regularised Random Forest (RRF) in gene selection. Despite its recentness, Boruta algorithm has shown its stability and effectiveness from the result consistency as compared to the existing RF method.

In Boruta, the features are extended with shadow. Shadow is known as artificial features that are created by permuting the order of values in the original data. These artificial features are used to gather the shadows' importance scores to judge the significance of the scores obtained by the actual features. The steps of the Boruta algorithm are as follows Kursa, (2014):

- Step 1: Duplicate the acoustical feature matrix into shadow.
- Step 2: Shuffle values in each shadow randomly.
- Step 3. Run Random Forest on the (shuffled features) and compute the Z scores.
- Step 4. Find the Maximum Z-score among Shadow Attribute (MZSA).
- Step 5. Run Random Forest on original features.

- Step 6: Assign each original feature a hit if feature Z-score $>$ MZSA.
- Step 7: If Z-score \leq MZSA, perform two-side equality test against MZSA.
- Step 8: If Z-score $<$ MZSA significantly, drop feature as unimportant.
- Step 9: If Z-score $>$ MZSA significantly, keep feature as important.
- Step 10: Repeat from step 5 until all importance is determined for all features.

4.5 SUMMARY

In this chapter, a new, and improved local maxima of speech energy (LM-E) is introduced as an acoustical feature to represent the expressive intonation of elongation. A ruled-based threshold calculation is proposed to determine the local maxima of the speech energy. Other than LM-E, eight well-established acoustical features are described and ranked based on its importance using Random Forest Boruta algorithm. In Chapter Five, the acoustical features are modeled using a statistical classifier to determine the best feature(s) that are able to discriminate Malay language filled pause and elongation.

CHAPTER FIVE

PATTERN CLASSIFICATION FOR CONSTRUCTION OF DISCRIMINATIVE MODEL

5.1 INTRODUCTION

This chapter presents the pattern classification to develop a model for discriminative classification of filled pause and elongation. In this chapter, two types of pattern classification are conducted. The first pattern classification is using single input while the second one is conducted using multi-input to the classifier. The main aim of this chapter is to determine how the acoustical features pattern should be modeled to best classify Malay language filled pause and elongation. The previous features used in classification were unable to classify filled pause and elongation since the existence of a model for Malay filled pause and elongation classification is non-existence. Design of the probability density estimation and the single feature classification using Bayes are then presented and discussed. The classification performance evaluations are done using recall, precision, F-measure and accuracy.

5.2 BAYES CLASSIFICATION PROCESS

The Bayes classification process is illustrated as in Figure 5.1. In general, the single pattern classification process is divided into two stages; the conditional probability density estimation and Bayes theorem computation. Each acoustical feature values that is gathered from the feature extraction stage in Chapter Four goes through individual conditional probability density estimation in obtaining and understanding the distribution of the underlying values. Based on the value of the conditional probability density, the Bayes theorem classifies the disfluencies into filled pause or elongation by choosing the class with higher conditional probability.

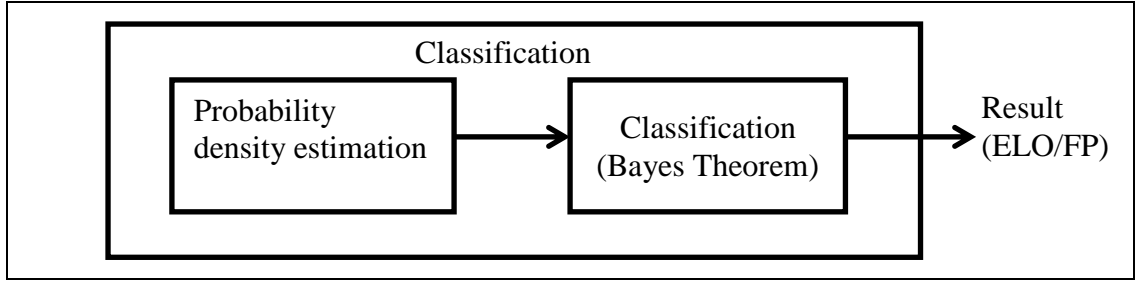


Figure 5.1: Single classification process

The process of single feature classification is described as follows:

- i. The classifier learns the conditional probability from the training data of the attributes X (acoustical feature values) given the class label, C (FP or ELO).
- ii. The classification is performed by applying Bayes rules to compute the probability of C given the particular feature of X .
- iii. The class of the feature X is predicted by the highest posterior probability.

Let x be a specific feature with assigned values of $x_1, x_2, x_3 \dots x_n$ and C is the class with assigned values of class variables of $C_1, C_2, C_3 \dots C_n$. The Bayes classifier enables the computation of the posterior probability $P(C = c_k | X = x)$ for each possible class c_k using Bayes theorem (Dougherty et al., 1995). The class label of the disfluency is determined by using Bayes theorem as in equation (5.1) (Papoulis, 1994).

$$C = \operatorname{argmax} p(C_i | x) \quad (5.1)$$

where,

$$p(C_i | x) = \frac{p(x | C_i)p(C_i)}{\sum_{i=1}^I p(x | C_i)p(C_i)} \quad (5.2)$$

and,

$$p(x, C_i) = p(x | C_i)p(C_i) \quad (5.3)$$

where,

c = class of the disfluencies ($c = \text{FP}$ for filled pause , $c = \text{ELO}$ for elongation)

x = acoustical feature

$p(c)$ = prior probability

$p(x/c)$ = conditional probability

$p(c/x)$ = posterior probability

The prior probability of 0.5 for each class is set equally since the number of filled pause and elongation is distributed equivalently.

5.2.1 Conditional Probability Density Estimation

In this research, prior to deciding the suitable probability density function, the data distribution is observed using histogram of each acoustical feature. Histogram is among the simplest method used to represent data distribution. However, each feature of elongations and filled pauses are not normally distributed as shown in Figure 5.2 and Figure 5.3. Therefore, instead of using Gaussian probability density estimation, this research employs Kernel Density Estimation (KDE).

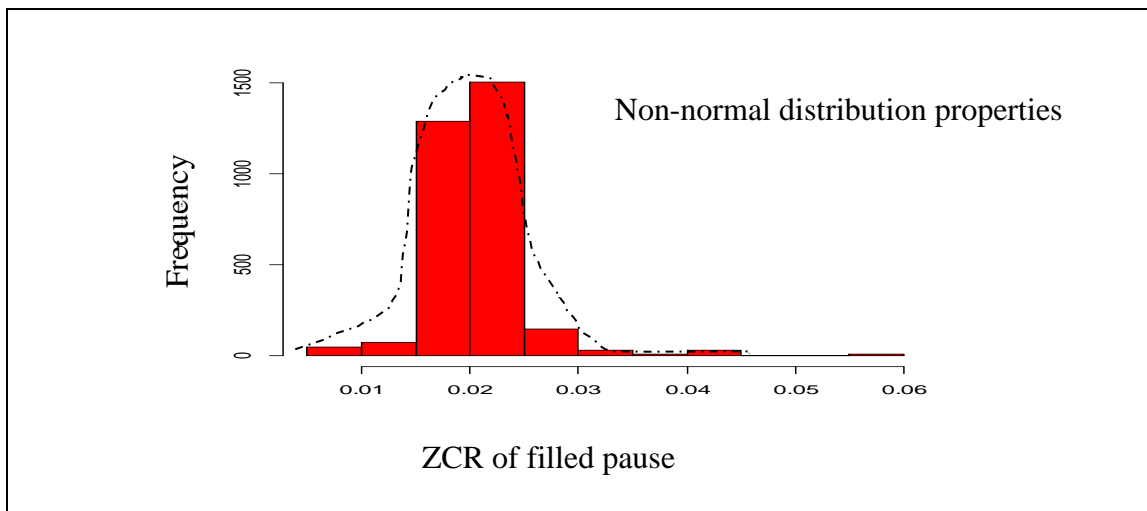


Figure 5.2: ZCR histogram for filled pauses

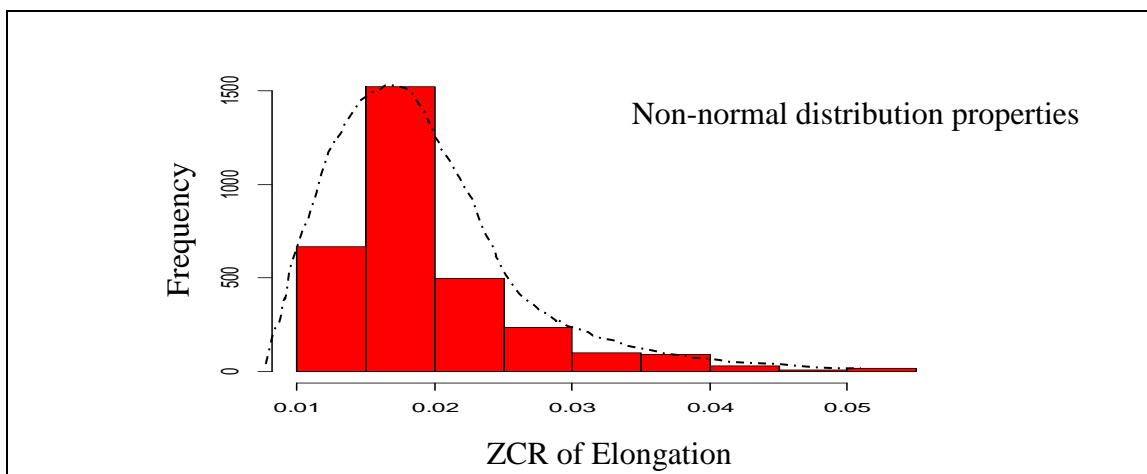


Figure 5.3: ZCR histogram for elongations

5.2.1.1 Kernel Density Function using Silverman's bandwidth

As stated in section 2.7.1.1 Silverman's rule of Thumb is chosen as bandwidth selection for KDE. Therefore, Kernel density estimation function utilizing Silverman's bandwidth selection is constructed and used to generate the conditional probability set of rules for each acoustical feature. For example, for ZCR feature's the KDE is represented as in Figure 5.4. The estimated Kernel density is represented in y-axis, while the x-axis represents the ZCR value of filled pause and elongation. It is observable that the Kernel density of elongation's ZCR which is in the range of (0-18) is higher compared to filled pause's ZCR. As mentioned earlier, Bayes theorem classification is made based on the highest conditional probability to classify the disfluencies as shown in equation (5.1). Therefore, if ZCR value of elongation falls within range (0-18) is correctly classified into elongation class. On the other hand, any ZCR of filled pause which is in the range of (18 to 40) is correctly classified into filled pause class as the density is higher compared to elongation.

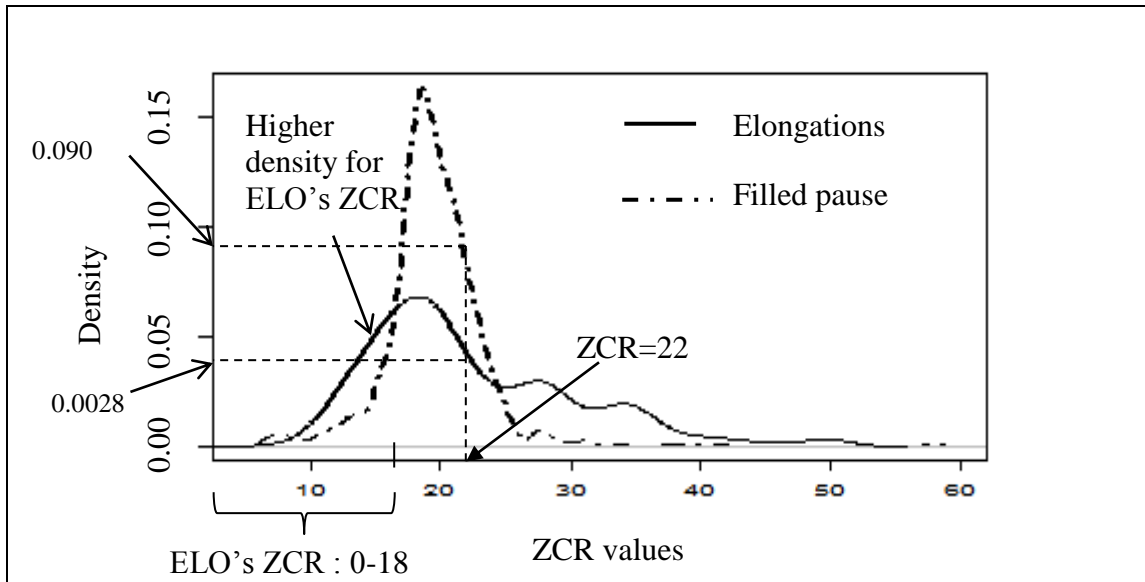


Figure 5.4: ZCR Kernel density plot for filled pauses and elongations

Kernel density estimation is used in the classification process. To show the process of classification using KDE and Bayes theorem, the ZCR is chosen as an example by referring to Figure 5.4. The example of how ZCR classification is done is shown as follows:

Condition 1: Disfluency with ZCR value of 22

$P(C_{FP})$ = Prior probability of FP : 0.5

$P(C_{ELO})$ = Prior probability of ELO: 0.5

$P(X_{ZCR}/C_{ELO})$ = Conditional probability of elongation's KDE: 0.0028

$P(X_{ZCR}/C_{FP})$ = Conditional probability of filled pause's KDE: 0.090

$P(C_{FP}/X_{ZCR})$ = Posterior probability for ZCR in FP class:

$$p(C_i | x) = \frac{p(x | C_i)p(C_i)}{\sum_{i=1}^I p(x | C_i)p(C_i)} \quad , C_i = C_{FP}$$

$$\frac{0.5 \times 0.090}{(0.5 \times 0.090) + (0.5 \times 0.0028)} = 0.97$$

$P(C_{ELO}/X_{ZCR})$ =Posterior probability for ZCR in ELO class:

$$p(C_i | x) = \frac{p(x | C_i)p(C_i)}{\sum_{i=1}^I p(x | C_i)p(C_i)} \quad , C_i = C_{ELO}$$

$$\frac{0.5 \times 0.0028}{(0.5 \times 0.090) + (0.5 \times 0.0028)} = 0.03$$

Therefore, the speech segment is classified as filled pause since it scores on the highest posterior probability at 0.97 compared to elongation at 0.03.

5.3 ACOUSTICAL RULES PARAMETER SELECTION FOR FILLED PAUSE AND ELONGATION

The third contribution of this research is presented in this section. Based on the previous conditional probability estimation by KDE on each acoustical feature, a set of acoustical rules are proposed in this research. An example of a feature (i.e. LM-E) is chosen to illustrate the rules parameter selection process. The KDE of LM-E is shown in Figure 5.5.

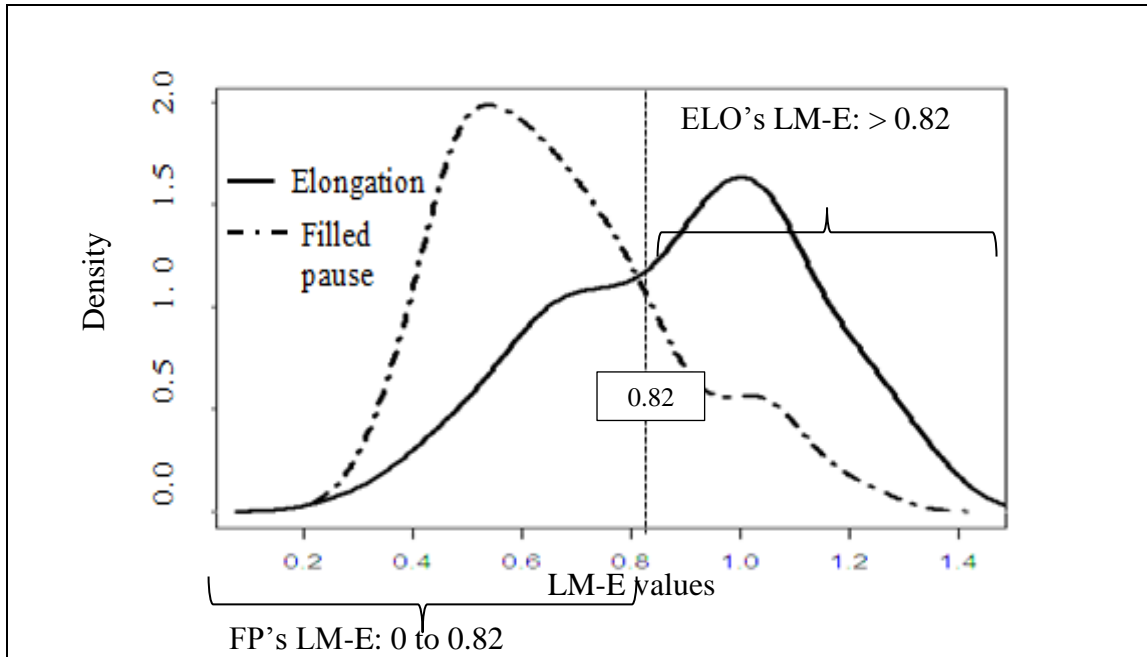


Figure 5.5: KDE of LM-E for filled pauses and elongations

From the figure, it can be seen that the density of LM-E for filled pause class is higher in the range of 0 to 0.82, while the elongation's KDE of LME is higher in the range of 0.82 to > 1.6. Therefore, the rules set for LM-E are:

LM-E: If $0 < \text{LM-E} \leq 0.82$ Then speech segment = Filled pause
 Else speech segment = Elongation

This mechanism is applied on each acoustical feature to come out with sets of acoustical feature parameter rules. The acoustical rules are presented as in Table 5.1. This probability value is used in the posterior probability calculation for the disfluencies classification by the Bayes theorem (refer to Section 5.2.1).

Table 5.1:
Proposed acoustical rules for filled pause and elongation classification

Acoustical feature	Filled pause	Elongation
ZCR	$18 < \text{ZCR} < 25$	$0 < \text{ZCR} < 18$ or $25 < \text{ZCR} < 60$
LM-E	$0 < \text{LM-E} < 0.82$	$\text{LM-E} > 0.82$
FF1	$0 < \text{FF1} < 25$	$25 < \text{FF1} < 150$
FF2	$0 < \text{FF2} < 100$	$100 < \text{FF2} < 970$
FF3	$0 < \text{FF3} < 120$	$120 < \text{FF3} < 620$
FF4	$0 < \text{FF4} < 150$	$150 < \text{FF4} < 820$
STE	$0 < \text{STE} < 200$	$200 < \text{STE} < 620$
F0	$0 < \text{F0} < 12$ or $37.5 < \text{F0} < 125$	$12 < \text{F0} < 38$ or $125 < \text{F0} < 200$
MFCC	$0.75 < \text{MFCC} < 1.22$	$0 < \text{MFCC} < 0.75$

The pseudo codes of the proposed acoustical rules for each feature are shown in Figure 5.6.

```
Pseudo Code:
Case: Feature
ZCR: If  $18 < ZCR < 25$  Then speech segment = Filled pause
      Else speech Segment = Elongation
LM-E: If  $0 < LM-E < 0.82$  Then speech segment = Filled pause
      Else speech segment = Elongation
FF1: If  $0 < FF1 < 25$  Then speech segment = Filled pause
      Else speech segment = Elongation
FF2: If  $0 < FF2 < 100$  Then speech segment = Filled pause
      Else speech segment = Elongation
FF3: If  $0 < FF3 < 120$  Then speech segment = Filled pause
      Else speech segment = Elongation
FF4: If  $0 < FF4 < 150$  Then speech segment = Filled pause
      Else speech segment = Elongation
F0: If  $0 < F0 < 12$  or  $37.5 < F0 < 125$  Then speech segment = Filled pause
      Else speech segment = Elongation
MFCC: If  $MFCC > 0.75$  Then speech segment = Filled pause
      Else speech segment = Elongation
STE: If  $STE > 0.75$  Then speech segment = Filled pause
      Else speech segment = Elongation
End case
```

Figure 5.6: Acoustical rules pseudo code

5.4 CROSS VALIDATION

In a supervised classification, a training set of examples with labels, and a test set of examples with unknown labels are usually provided. The main reason is to classify the test set to either filled pause or elongation class. Training and testing are common processes in classification or pattern recognition to validate the classifier and is a usual process in observing a classifier's accuracy. There are different ways of

validating classifier found in previous researches. Cross validation (CV) is the most common and recently used (Elkan, 2012). There are several techniques applied in CV such as leave one-out and fold-CV. In Tanaka et al., (2014), a total of 186 iterations is applied into leave one-out cross-validation in which each time of the experiment, 1 sequence of data is taken out as a test data while the rest is used for training. This process is repeated up to 186 times (Tanaka et al., 2014). However, this method is quite time consuming for a larger dataset. A large dataset that consists of 1076 samples has been applied with 10-fold CV in order to test the classifier's performance (Elkan, 2012). The study found that their classifier's performance is comparable with the previous work done by Bouckaert, (2004). In (Murakami & Mizuguchi, 2010), two stage of classifiers validation is done. The first stage is conducted by using standard training and testing data partition with different data division ratios while the second stage uses cross-validation.

This research chooses cross validation method to test the accuracy of the model. The number of fold chosen is 10 as it is the frequently used number while validating the developed model as seen in (Paja & Wrzesie, 2013; Womack, 2012; Soria et al., 2011; Loh, 2011; Rimer, 2007; Schuller et al., 2005). In 10-CV technique, nine folds are used to train classifier, and the one fold that is held out is then used to test the classifier. The process of dividing the data into 10-fold CV is as follows (Elkan, 2012):

Input: Training set S , integer constant K

Procedure:

Partition S into K equal-sized subset $S_1 \dots S_n$

for $i = 1$ to $i = k$

Let $T = S / S_i$

Run learning algorithm (Bayes classifier) with T as training set

Test the resulting classifier on S_i .

The process of 10-fold cross validation is illustrated as in Figure 5.7. The total data of filled pause and elongation are divided into 10 equivalent folds. This process is executed 10 times with different fold used as testing during each iteration.

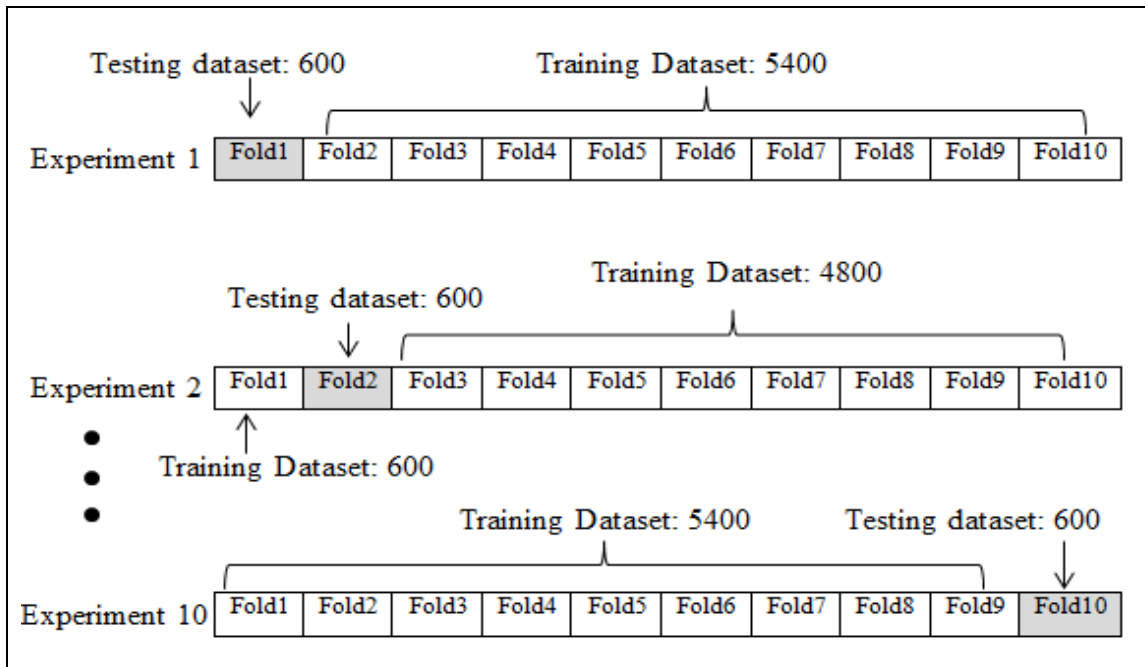


Figure 5.7: Ten-fold cross-validation process

5.6 DISCRIMINATIVE CLASSIFICATION MODEL CONSTRUCTION USING NAÏVE BAYES

The aim of this section is to produce a discriminative model for filled pause and elongation classification so that the extracted acoustical features can be combined and produce a higher classification performance. The discriminative classification modeling is executed by using feature based combination in Naïve Bayes algorithm. Generally, the classification involves three stages of different experiments to evaluate the performance of the discriminative model. Each discriminative classification was conducted using 10-fold cross validation as the previous single feature classification.

In this research, the Naïve Bayes product rule is employed to combine the acoustical features in order to get the best discriminative classification model for filled pause and elongation. Two acoustical features (LM-E and ZCR) that are proven significant for Malay filled pause and elongation through the feature ranking and single feature classification are considered in this feature combination classification. Naïve Bayes is the simplest Bayesian network classifier inclusive of a node C which represents the class and attributes $x_n = (x_1, x_2, \dots, x_n)$ as features as illustrated in Figure 5.8.

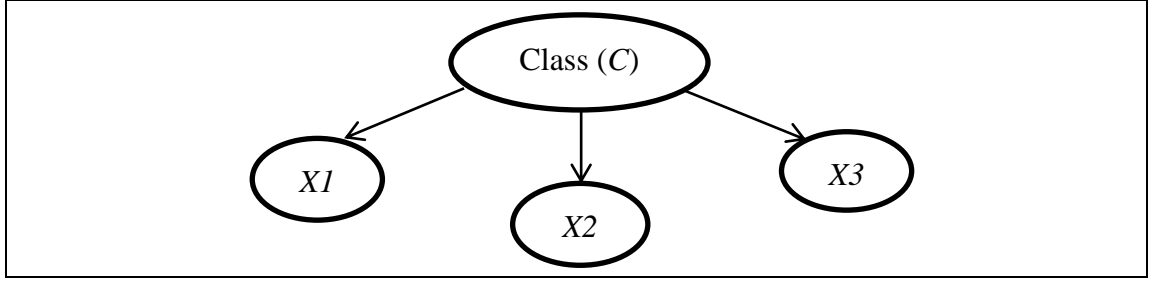


Figure 5.8: Naïve Bayes illustration

In relation to this research, the acoustical features that are extracted are independent of each other. Although some of the features such as formant frequency 1-4 (FF1, FF2, FF3, and FF4) are in the same class of feature, each formant is differentiated by the levels qualifying each of them as individual feature. The process of discriminative classification model construction is as follows:

- i. Expand equation (5.3) from one to multiple conditional probabilities as equation (5.4).

$$P(X_n | C_j) = \frac{P(C_j)P(X_1 | C_j) * P(X_2 | C_j) * ... * P(X_n | C_j)}{\sum_{i=1}^I P(X | C_j)P(C_j)} \quad (5.4)$$

where,

$P(X_1 | C_j) * P(X_2 | C_j) * P(X_n | C_j)$ is the probability of class C_j generating the observed value for first acoustical feature X_1 , multiplied by the probability of class C_j generating the observed value for second acoustical feature X_2 followed by the multiplication with the probability of class C_j generating the observed value for n acoustical feature X_n .

- ii. The classifier learns the multiple conditional probabilities produced by different acoustical features.
- iii. Perform Bayes rules as equation (5.1) to get the maximum probability for the classification of filled pause and elongation.
- iv. Assign the class of the disfluency based on the highest probability.

The Naïve Bayes classification of the discriminative model (LM-E+ZCR) is illustrated as in Figure 5.9. In Naïve Bayes classification, each acoustical feature acts as a sub classifier and went through a preliminary classification. Each of the

preliminary classification produced conditional probability that is then combined with each other to produce the final posterior probability.

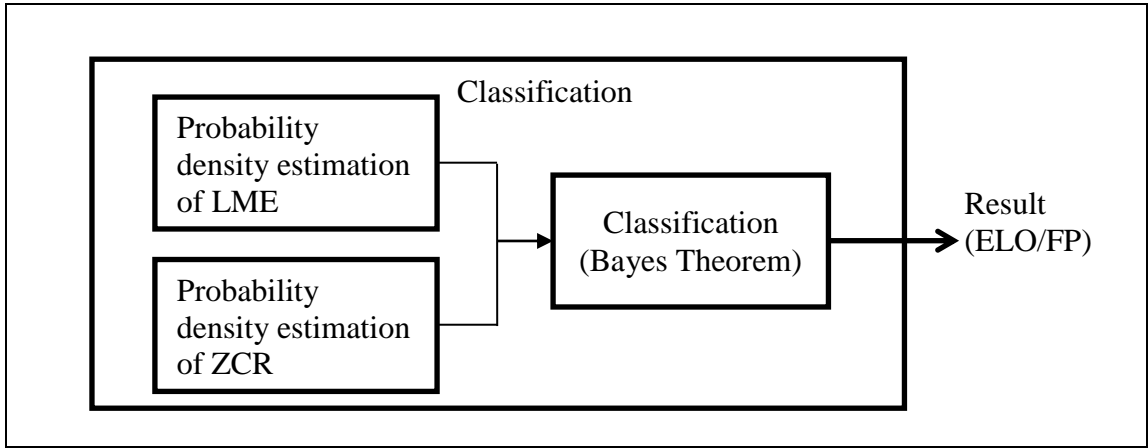


Figure 5.9: Illustration of Naïve Bayes fusion classification

The discriminative model is defined as in equation (5.5).

$$P(X_n | C_j) = \frac{P(C) \cdot (P(X_{LM-E} | C_{FP/ELO}) * P(X_{ZCR} | C_{FP/ELO}))}{P(X_{LM-E})P(C_{FP/ELO}) + P(X_{ZCR})P(C_{FP/ELO})} \quad (5.5)$$

where;

n = LM-E or ZCR

C_j = FP or ELO

$P(X_{LM-E} | C_{FP/ELO})$ = Conditional probability of LM-E generating class filled pause or elongation

$P(X_{ZCR} | C_{FP/ELO})$ = Conditional probability of ZCR generating class filled pause or elongation

In the discriminative classification experiment, three types of experiments are conducted to compare the impact of LM-E feature when combined with the other techniques. It is also conducted to observe the classifier performance when LM-E feature is not involved in the classification and to observe whether the number of acoustical features used is more important compared to relevant feature used in the classification. The description of each experiment is as follows:

Experiment 01:

The LM-E and all standard acoustical feature are used in the classification. This experiment is conducted to identify which acoustical feature is best fused with LM-E to achieve the highest classification performance of filled pause and elongation.

Experiment 02:

The ZCR is combined with each standard acoustical feature. This experiment is executed to observe the fusion performance of ZCR and other standard acoustical features.

Experiment 03:

In the last experiment, the fusion of all standard acoustical features (ZCR+F0+MFCC+FF+STE) is done. The result of Experiment 03 is then compared with the discriminative model performance.

5.7 EVALUATION METHOD

The evaluation of each experiment is conducted by computing the recall, precision, F-measure, and accuracy. The evaluation is done to assess the classification model of filled pause and elongation. Calculations of these measures are as follows:

$$\begin{aligned} \text{Recall} & \quad (5.6) \\ &= \frac{\text{Number of relevant filled pause or elongation correctly classified}}{\text{All relevant filled pause or elongation}} \end{aligned}$$

$$\begin{aligned} \text{Precision} & \quad (5.7) \\ &= \frac{\text{Number of relevant filled pause or elongation correctly classified}}{\text{All classified filled pause or elongation}} \end{aligned}$$

$$F - \text{measure} = \left(\frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \right) \times 2 \quad (5.8)$$

$$\text{Accuracy} = \frac{\text{Total of correctly classified filled pause and elongations}}{\text{total filled pause and elongations}} \quad (5.9)$$

5.8 SUMMARY

This chapter presented the single feature classification model using Kernel Density Estimation (KDE) and Bayes theorem. Silverman's rule of thumb is also identified as KDE's bandwidth selection. A new set of acoustical rules to be incorporated in the discriminative classification model to classify Malay language filled pause and elongation is proposed. This chapter concludes by fulfilling Objective 3 of this thesis that is to model the discriminative properties of filled pause and elongation acoustical feature and to assess the accuracy of the proposed model.

CHAPTER SIX

RESULTS AND DISCUSSION

6.1 INTRODUCTION

This chapter presents the results gathered from each stages of methodology conducted in this thesis. The purpose of this chapter is to present the result flow of this thesis that lead from one stage to the other. The results are presented in five sections starting from the results of voice activity detection, Random Forest-based feature ranking, classification of single pattern and the performance of discriminative model of filled pause and elongation classification. An additional section is added to test the robustness of the discriminative classification model using English language datasets.

6.2 VOICE ACTIVITY DETECTION RESULTS FOR NOISE REMOVAL

VAD in this research is conducted using energy and zero crossing rates as threshold. To evaluate the performance, formant frequency acoustical feature is chosen as a benchmarked measurement because it is one of the established acoustical features that is being used in filled pause researches. Therefore, for the evaluation of VAD performance in this research the formant frequency is used. The details of the formant frequency extraction are discussed in Chapter Five (i.e. Acoustical Features Extraction). According to (Rabiner & Sambur, 1975) and (Bachu et al., 2008), the formant frequency's standard deviation of normal words is higher compared to filled pause. This research takes FP10.wav as an example of speech to be evaluated. The purpose is to observe the formant characteristics on each VAD method. Through the observation, by using energy-based VAD, the desired voiced region of the speech utterances are not exactly detected as can be seen in Figure 6.1.

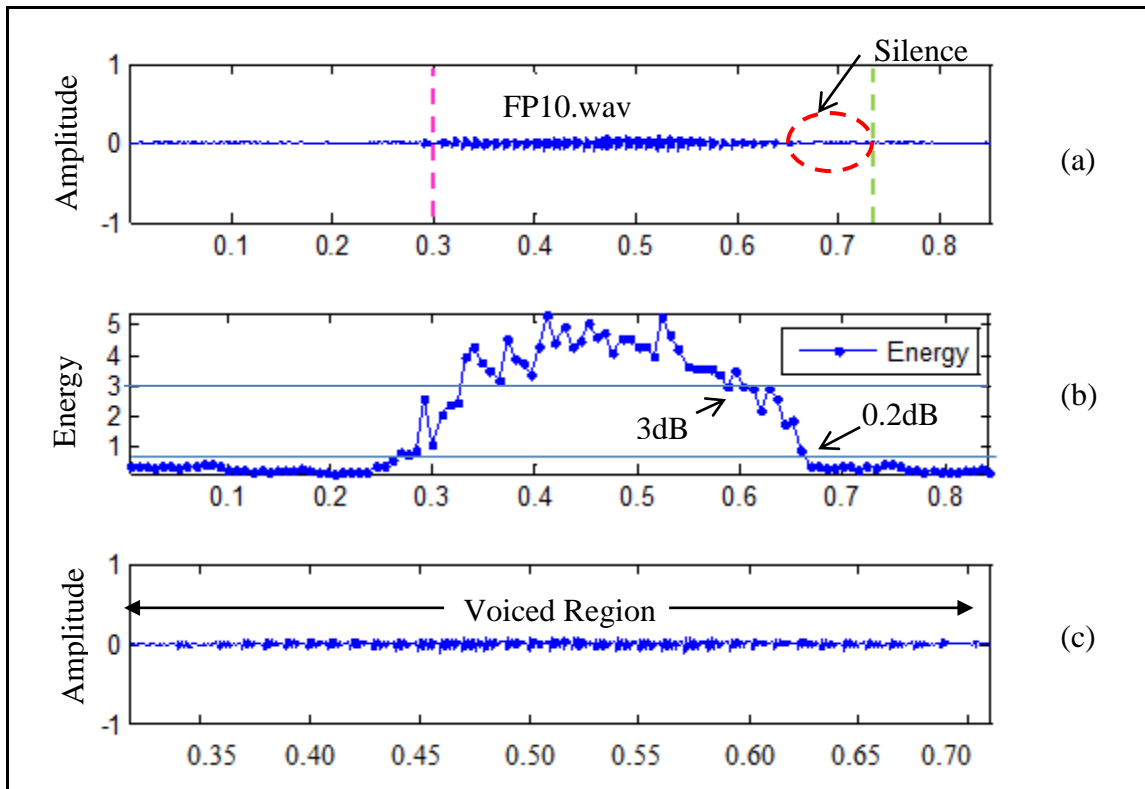


Figure 6.1: Voice region detected by energy-based VAD

From Figure 6.1(a), it is observable that the voiced region is inaccurately detected since part of the silence is included in the voiced region. Referring to the time interval between 0.65sec to 0.73sec (refer to the dotted red circle), the energy-based VAD algorithm has detected the silence region as voiced. When silence is included, the speech volume indicated by amplitude dropped from 3dB at 0.6sec to 0.2dB at 0.65sec. The standard deviation of formant frequencies of the speech of is calculated and shown in Table 6.1. The abrupt changes of formant frequencies value causes formant frequencies data to become unstable and thus, producing larger standard deviation. A comparison is made to show the standard deviation before and after removal of the silence region. The removal is done manually for the purpose of analysis only. The result is shown as in Table 6.1. From the results, the manual removal of unvoiced speech segment significantly reduced the standard deviation of the formant frequencies. It indicates that the VAD is important in filled pause and elongation classification to ensure accurate voiced speech segment is acquired.

Table 6.1:
Effects on formant frequencies standard deviation before and after unwanted speech interval removal for energy-based VAD

Interval removal	σ FF1	σ FF2	σ FF3	σ FF4
Before	53.41	112.92	94.00	117.55
After	13.00	61.71	24.08	68.51

By using the same speech sample, combination of energy + ZCR-based (EZCR) VAD; and energy + HOD-based (EHOD)VAD are analyzed. Results of energy + ZCR-based VAD are presented in Figure 6.2 and outcome of energy + HOD-based VAD are illustrated in Figure 6.3.

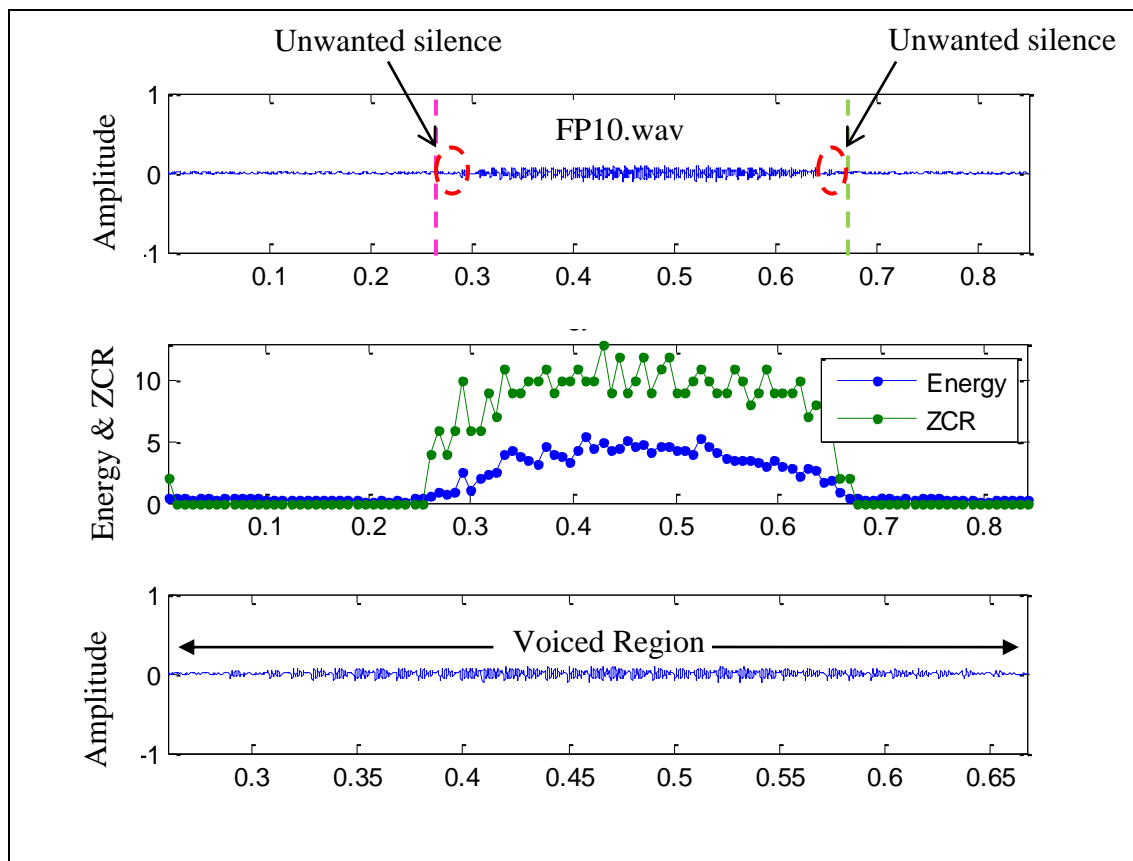


Figure 6.2: Voice region detected by energy + ZCR-based VAD

From the results shown in Figure 6.2, the voiced region detection by energy and ZCR-based method for FP10.wav is more accurate compared to the energy-based method. It can be seen that less silence region is detected by the EZCR at the interval of 0.27s to 0.3s and 0.64s to 0.65s. The statistical observation on the formant frequencies standard deviation is presented in Table 6.2. The result clearly showed that the formant frequencies standard deviation reduced after the detected silence

region of the speech is removed. From Table 6.2, for example, the standard deviation of FF1 is reduced from 53.86 to 13.32 resulting in 75.31% reduction; while for FF2, the standard deviation of the formant frequency is reduced at 47.96% from 109.19 to 56.82. This phenomena is due to the ability of combined threshold method of energy and ZCR to remove the silence region.

Table 6.2:
Effects on formant frequencies standard deviation before and after unvoiced removal for energy and ZCR based-VAD

Unvoiced removal	σ FF1	σ FF2	σ FF3	σ FF4
Before	53.87	109.19	89.55	102.99
After	13.32	56.83	21.98	64.18

The energy + HOD-based VAD result is also being analyzed by using the same procedure and same speech sample. The result of energy + HOD-based VAD is shown in Figure 6.3. From the figure, it is clearly seen that the voiced region is detected accurately compared to the previous energy-based method where no silence region is included in the detected voiced segment of the speech. The voiced region detected for FP10.wav started at 0.31s to 0.61s.

The statistical observation on the formant frequencies standard deviation is presented in Table 6.3. The results clearly show that the formant frequencies standard deviation of energy + HOD - based VAD is lower by 48% for σ FF1, 30.5% for σ FF2, 71.12% for σ FF3 and 28.10% for σ FF4 compared to energy + ZCR-based methods.

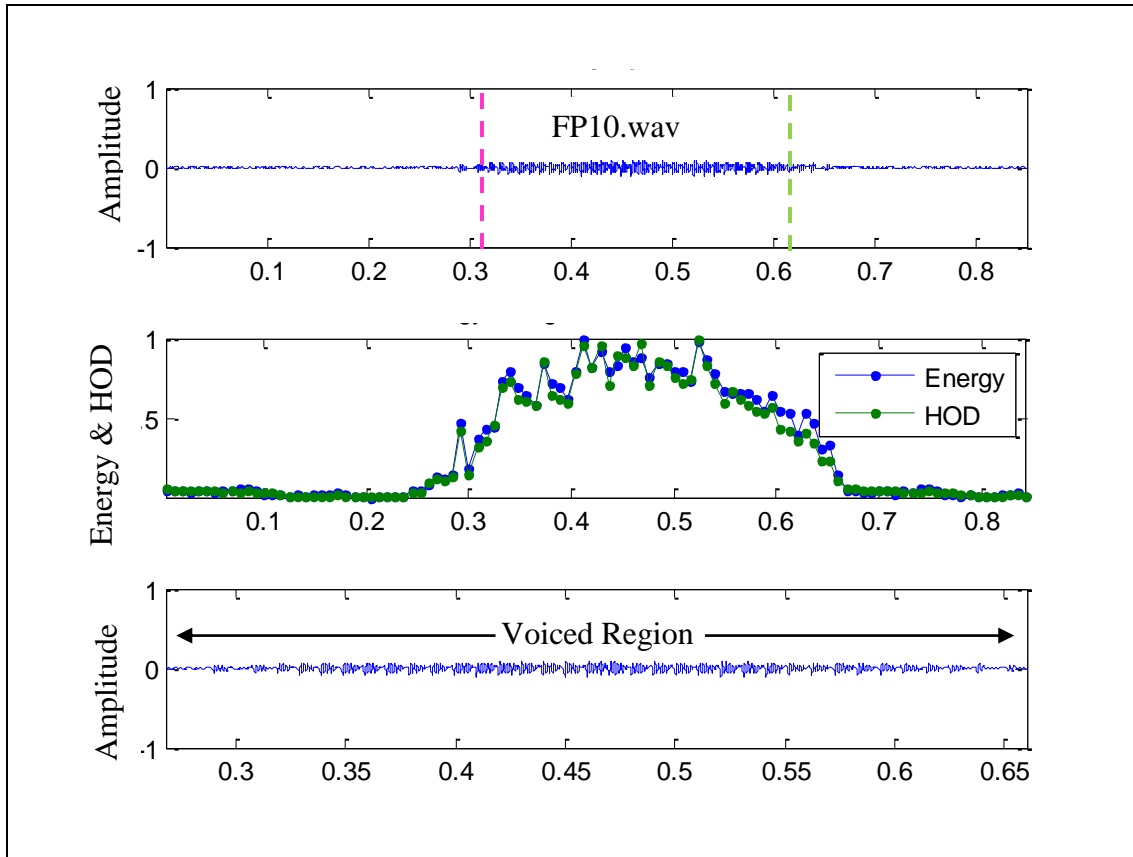


Figure 6.3: Voice region detected by energy-HOD based VAD

Table 6.3:

Formant frequencies standard deviation comparison

VAD methods	σ FF1	σ FF2	σ FF3	σ FF4	Average
Energy+HOD	28.00	76.11	25.86	74.05	51.01
Energy+ZCR	53.87	109.19	89.55	102.99	88.90
% comparison Energy+HOD and Energy+ZCR	48.00	30.50	71.12	28.10	42.62
Energy	53.41	112.92	94.00	117.55	94.47

The formant frequencies standard deviation for each datasets is calculated to get the overall VAD measurement. The mean of the overall formant frequencies standard deviation is calculated and shown as in Table 6.4. For description of each VAD dataset, please refer to Table 3.6. From the results, it is obviously seen that VAD_03 dataset of filled pause and VAD_06 dataset of elongation achieved the lowest average of formant frequencies standard deviation. The highest average of formant frequencies standard deviation is seen in VAD_01 and VAD_04 datasets, that are produced using

energy-based VAD. Thus, VAD_03 and VAD_06 datasets produced using energy + HOD-based VAD are used in the acoustical feature extraction stage.

Table 6.4:
Mean of formant frequencies standard deviation of each datasets

Datasets	$\mu\sigma$ Formant_1	$\mu\sigma$ Formant_2	$\mu\sigma$ Formant_3	$\mu\sigma$ Formant_4
VAD_01	42	156	209	199
VAD_02	40	149	193	186
VAD_03	<u>38</u>	<u>148</u>	<u>194</u>	<u>175</u>
VAD_04	48	194	210	183
VAD_05	43	189	215	183
VAD_06	<u>32</u>	<u>175</u>	<u>203</u>	<u>177</u>

6.3 FEATURE RANKING FOR PAUSE AND ELONGATION REPRESENTATION

After voice activity detection, nine acoustical features as discussed in Chapter Four are extracted from VAD_03 (i.e. filled pause) and VAD_06 (i.e. elongation) datasets. Feature ranking is then applied to all acoustical features to measure their rank of importance in representing Malay language filled pause and elongation.

The results of feature ranking are tabulated in Table 6.5. This table tabulates the results gathered by using random seed of 500. This table contains different values of Z-score for each feature: mean, median, minimum and maximum. The last column contains decision about confirmation or rejection of features indicating the status of the feature i.e. important or not important.

Among the acoustical features, ZCR and LM-E are denoted as among the highest importance measure of 42.8 and 41.5 mean Z-score and followed by FF2 with 50% difference of mean Z-score of 24.3. The median, min and max Z-score of ZCR and LM-E also denote the highest scores among the other acoustical features. The descending orders of the feature importance are then followed by FF1, STE, FF3, MFCC, F0, and FF4.

Feature ranking using Boruta algorithm feature showed that ZCR and LM-E achieved the highest mean Z-score, respectively. The following ranks are STE, FF1, FF3, MFCC, F0 and FF4. The results showed that ZCR and LM-E are the most suitable acoustical features that can be used to represent Malay language filled pause and elongation.

Table 6.5:
The Z-score for each acoustical features and decision about its importance

Acoustical Features	MeanZ	MedianZ	MinZ	MaxZ	Decision
FF4	7.8	7.8	4.6	9.6	Confirmed
F0	10.1	10.1	8.1	11.8	Confirmed
MFCC	15	14.9	12.9	17	Confirmed
FF3	16.4	16.2	15.1	18.2	Confirmed
STE	23.3	23.3	21.5	25.3	Confirmed
FF1	20.7	20.6	18.9	22.6	Confirmed
FF2	24.3	24.5	22.5	26.4	Confirmed
LM-E	41.5	41.4	39.2	44.8	Confirmed
ZCR	42.8	42.7	40.2	46.1	Confirmed

6.4 SINGLE FEATURE BAYES CLASSIFICATION PERFORMANCE EVALUATION

Single-feature classification on each of nine acoustical features are conducted using Silverman's Kernel Density Estimation and Bayes Theorem. The purpose of this classification is to identify the most contributing features for discriminating filled pause and elongation into two separate classes. Results of single-feature classification are then compared to feature ranking to confirm the ranks of importance.

The single-feature classification performance are measured using precision, recall, F-measure and accuracy. The precision and recall rate are needed to get the F-measure. The recall rate shows that the number of relevant filled pause or elongation that is successfully classified among the relevant filled pause or elongation. Precision shows the number of relevant filled pause or elongation that is successfully classified among all of the filled pause or elongation. On the other hand, F-measure is the harmonic mean between precision and recall rate. The accuracy shows the overall performance which denotes the number of filled pause or elongation that is successfully classified among all of the filled pause and elongation.

The overall results for each fold of the F-measure is presented in Table 6.6 and summarized in Figure 6.4. The precision and recall rate that are used in the F-measure calculation are presented in Table 6.7 and Table 6.8, which is then summarized in

Figure 6.5. Then, the accuracy of the each feature's performance is shown in Table 6.9 and summarized in Figure 6.6.

Among the entire features, LM-E and ZCR scored higher recall and precision rate at $> 68\%$ for both filled pause and elongation compared to the other acoustical features as can be seen in Figure 6.4. Overall, the highest F-measure for filled pause is achieved by ZCR at 79% followed by LM-E at 75%. LM-E scored the highest F-measure at 71% for elongations followed by ZCR at 63%. The LM-E achieved the highest F-measure of 71% for elongation compared to the other features. It shows that the proposed LM-E represents elongation better compared to the other acoustical features. As stated previously, the purpose of proposing LM-E is to represent elongation better by calculating the transition between the consonant and vowel (*refer Section 4.3*). The results of recall, precision and F-measure shows ZCR and LM-E are the top two ranks, which is comparable with the feature ranking that is previously discussed in section 6.3.

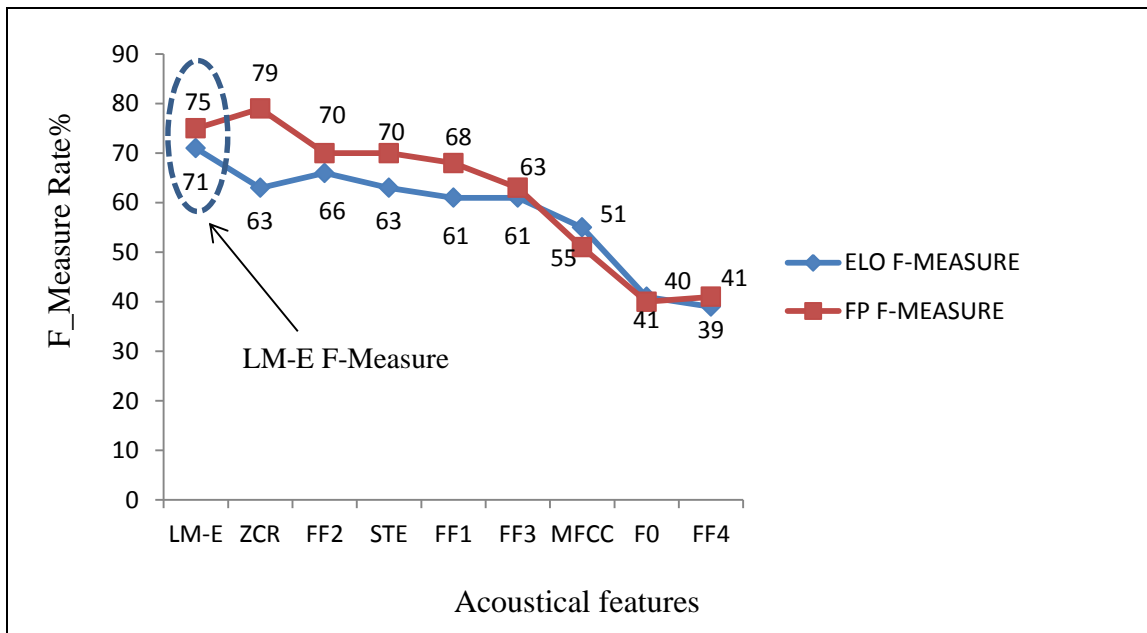


Figure 6.4: F-Measure of 10-fold CV

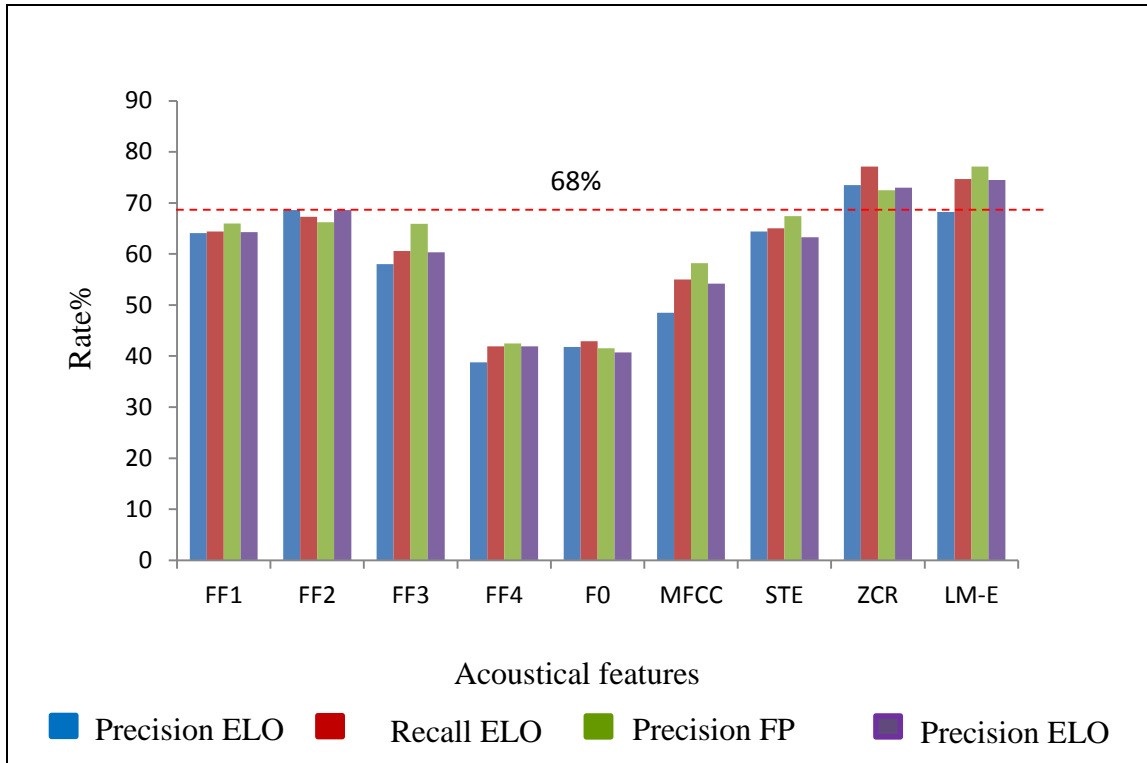


Figure 6.5: Recall and precision average of 10-fold CV

The overall correctness performance in discriminating filled pause is summarized using the average of the accuracy. It can be seen from Table 6.9 that both LM-E and ZCR achieved the best accuracy rate of 74% followed by FF2 at 68%. The accuracy difference between folds for each acoustical feature is represented in the standard deviation calculation.

Among the acoustical features, the highest standard deviation is seen in FF2 at 5.46 while the lowest is denoted in ZCR at 3.06. For the proposed LM-E, the accuracy differences between fold is considerably small which is only 3.89. This indicates that ZCR and LM-E is consistent in representing each filled pause and elongations. The lowest accuracy of the proposed LM-E is denoted at 68% as seen in the 7th fold. Most of the speech data of the 7th fold are from DR20080528 and DR20080828 datasets.

Table 6.6:

F-measure of 10 fold CV on single feature classification

Feature	Fold01		Fold02		Fold03		Fold04		Fold05		Fold06		Fold07		Fold08		Fold09		Fold10	
	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP
FF1	62	60	63	58	53	64	72	71	66	64	64	66	65	64	65	65	62	67	67	70
FF2	69	69	62	62	59	63	78	77	68	59	64	65	70	70	71	63	63	64	72	76
FF3	64	67	64	62	49	61	57	60	60	64	50	58	57	60	61	63	61	67	66	64
FF4	41	38	38	38	29	33	44	46	39	40	50	58	41	41	37	46	41	40	40	40
F0	40	40	39	37	30	36	45	52	48	33	44	39	43	41	40	45	47	41	43	43
MFCC	57	56	52	57	44	46	53	56	57	57	55	58	26	66	46	48	52	51	52	58
STE	62	63	70	63	63	62	68	65	63	65	62	68	55	63	60	60	69	74	72	68
ZCR	80	77	76	72	66	66	82	81	78	74	66	65	70	70	70	73	78	66	83	81
LM-E	69	76	68	76	66	73	75	76	67	74	77	83	70	65	70	75	72	74	78	82

Table 6.7:

Precision of 10 fold CV on single feature classification

Feature	Fold01		Fold02		Fold03		Fold04		Fold05		Fold06		Fold07		Fold08		Fold09		Fold10	
	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP
FF1	60	63	65	61	58	68	72	73	71	65	60	65	64	65	62	65	65	66	64	69
FF2	70	69	60	62	60	61	78	78	76	52	66	67	65	75	75	58	61	62	75	78
FF3	68	70	69	70	43	70	55	59	55	73	44	66	55	60	60	60	63	68	68	63
FF4	39	35	38	35	29	32	43	47	38	40	44	66	40	41	34	49	42	38	41	42
F0	39	40	38	35	30	38	40	55	53	30	45	38	40	42	40	47	50	43	43	47
MFCC	59	60	60	58	40	55	52	53	45	55	58	55	16	91	45	50	55	53	55	52
STE	60	63	72	63	68	65	69	68	65	64	60	73	50	69	55	60	70	77	75	72
ZCR	75	74	65	70	68	66	84	80	83	69	66	66	70	73	65	80	75	67	84	80
LM-E	66	81	64	81	61	79	71	80	65	75	83	78	70	63	62	68	64	80	76	86

Table 6.8:
Recall of 10 fold CV on single feature classification

Feature	Fold01		Fold02		Fold03		Fold04		Fold05		Fold06		Fold07		Fold08		Fold09		Fold10	
	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP	ELO	FP
FF1	65	58	62	55	49	61	73	70	62	63	68	67	66	63	69	66	60	68	70	72
FF2	69	70	64	63	59	65	79	77	62	69	63	64	75	66	67	70	65	67	70	75
FF3	61	65	59	55	57	54	60	61	65	57	57	52	60	61	63	66	60	67	64	65
FF4	43	41	39	42	30	35	46	45	41	40	57	52	43	41	40	43	40	42	40	38
F0	41	40	41	40	31	35	52	50	44	37	43	40	47	41	41	44	45	40	44	40
MFCC	55	53	46	56	50	40	55	59	78	60	53	61	65	52	48	46	50	50	50	65
STE	64	63	69	63	59	59	67	63	62	67	65	64	61	58	65	60	68	72	70	64
ZCR	85	81	92	74	64	66	81	83	73	79	67	65	70	68	75	67	81	65	83	82
LM-E	72	71	72	71	71	68	79	73	70	73	71	89	70	68	80	84	82	69	80	79

Table 6.9:
Accuracy of 10-folds cross validation on the single feature classification

Feature	Fold01	Fold02	Fold03	Fold04	Fold05	Fold06	Fold07	Fold08	Fold09	Fold10	Average	σ
LM-E	73	73	70	75	71	80	68	73	73	80	74	3.89
ZCR	72	75	72	75	71	72	75	69	80	75	74	3.06
FF2	70	62	61	78	64	65	70	67	64	74	68	5.46
STE	62	67	63	67	64	65	68	68	72	70	67	3.13
FF1	61	61	59	72	65	65	64	65	65	69	65	3.84
FF3	66	63	55	59	62	61	59	62	64	65	62	3.27
MFCC	57	55	45	55	57	57	54	47	52	55	53	4.22
F0	40	38	33	49	41	41	42	43	44	43	41	4.14
FF4	39	38	31	45	40	40	41	41	40	40	40	3.50

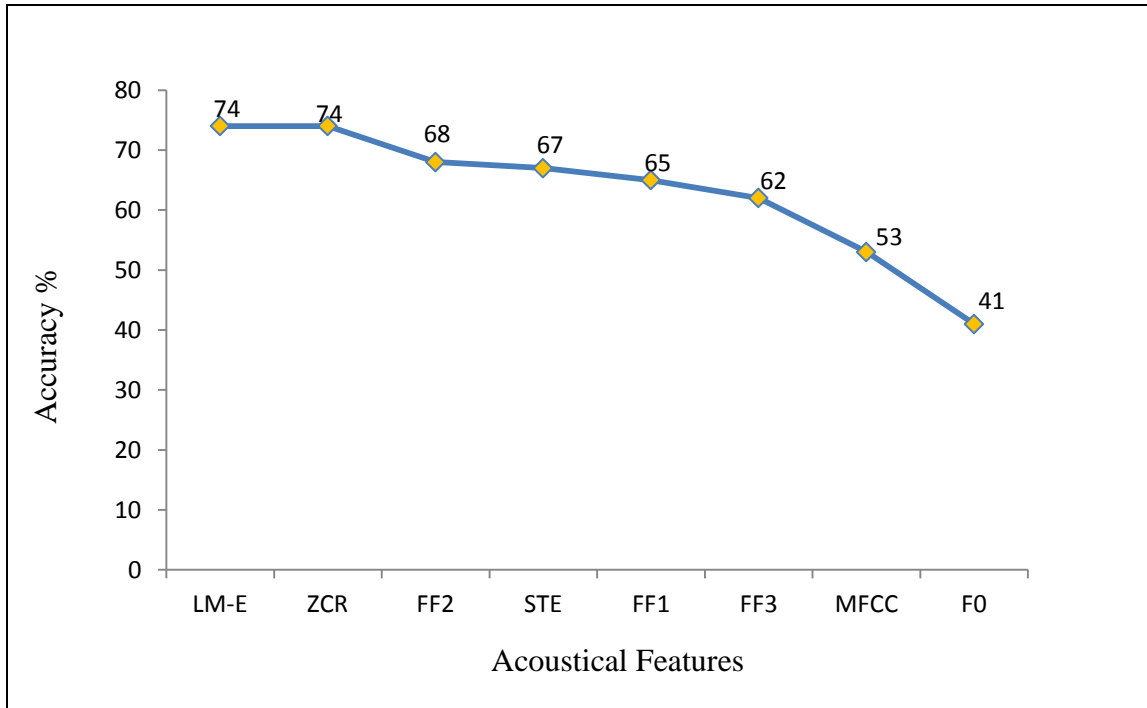


Figure 6.6: Average accuracy of each acoustical feature

A closer inspection in the 7th fold was done to find the reasons behind the lowest performance of the proposed LM-E. Upon observation on the 7th fold, it is discovered that there are too many filled pauses uttered with expressive intonation. Based on the observation on the datasets, the expressiveness in the filled pause utterance is produced when the speakers are in moody situation such as upset, angry, excited and happy. Some of the MPHD topic is discussed aggressively by the Malaysian Parliament members. The examples of filled pause (FP11.wav and F107.wav) that are unable to be correctly classified by LM-E are chosen and shown in Figure 6.7. The filled pauses are from the ‘eem’ type and are spoken by Speaker_01 and Speaker_02. The calculated LM-Es of the speech segment examples are also shown to be compared with the LM-E rules. Referring to the acoustical rule parameter (Section 5.3), in order for a disfluency to be detected as a filled pause; it should fulfill the rule of:

*“LM-E: If $0 < LM-E < 0.82$ Then speech segment = Filled pause
Else Speech segment = Elongation”*

However, the LM-E of both filled pause examples (0.9337 and 0.9828) are higher than 0.82. Since these values lie in the elongation interval of the LME’s KDE (refer to Figure 5.5), they are misclassified as elongation.

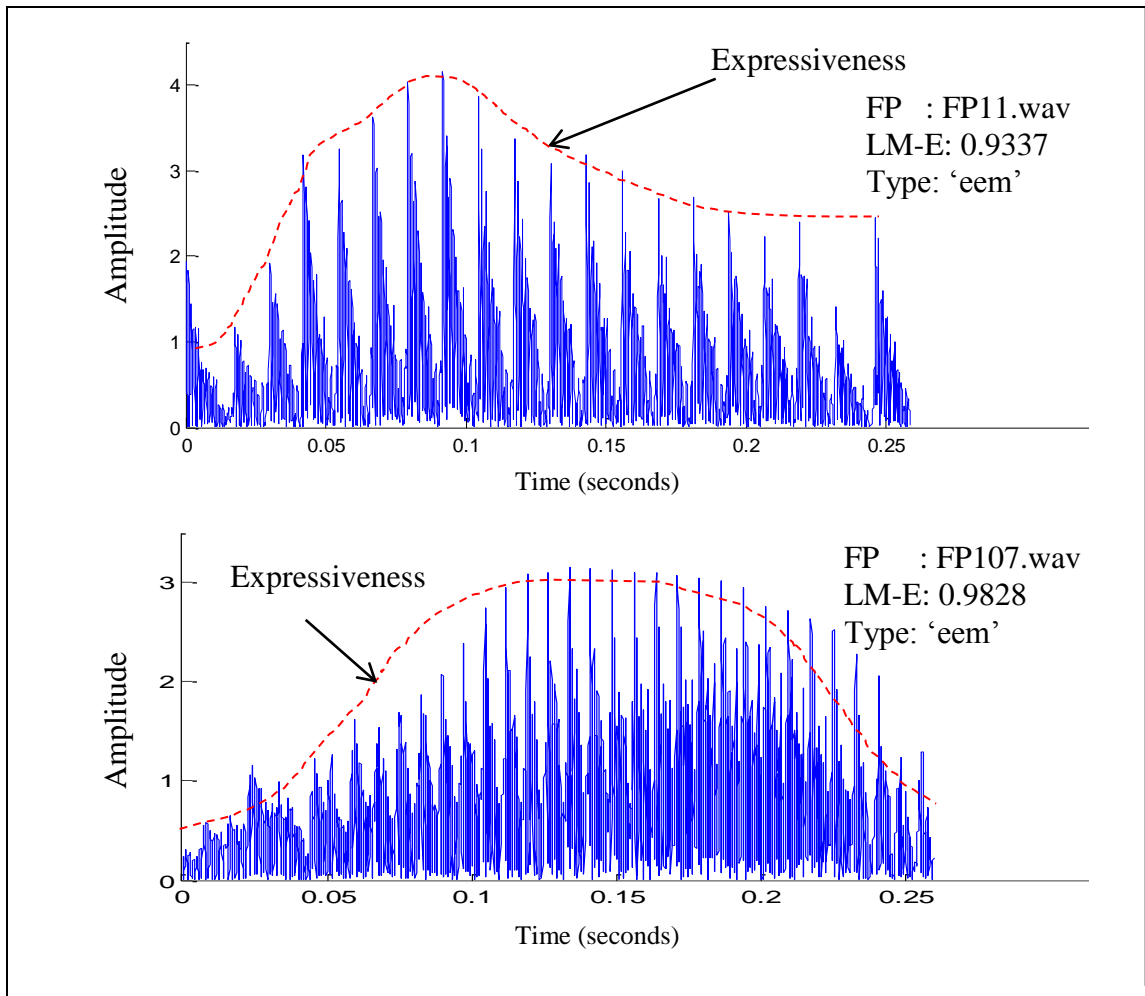


Figure 6.7: Example of misclassified filled pause

The examples of elongation which are misclassified as filled pause are shown in Figure 6.8. The examples are taken from DR20080828 dataset and were spoken by the same speaker. From the calculation, the LM-E of these examples is 0.684 for ELO06.wav for the syllable ‘ya’ and 0.378 for ELO07.wav of the syllable ‘wa’. Based on the LM-E rules, in order for a disfluency to be detected as an elongation; it should fulfill the rule of:

*“LM-E: If LM-E>0.82 Then speech segment = Elongation”
Else Speech segment = Filled pause”*

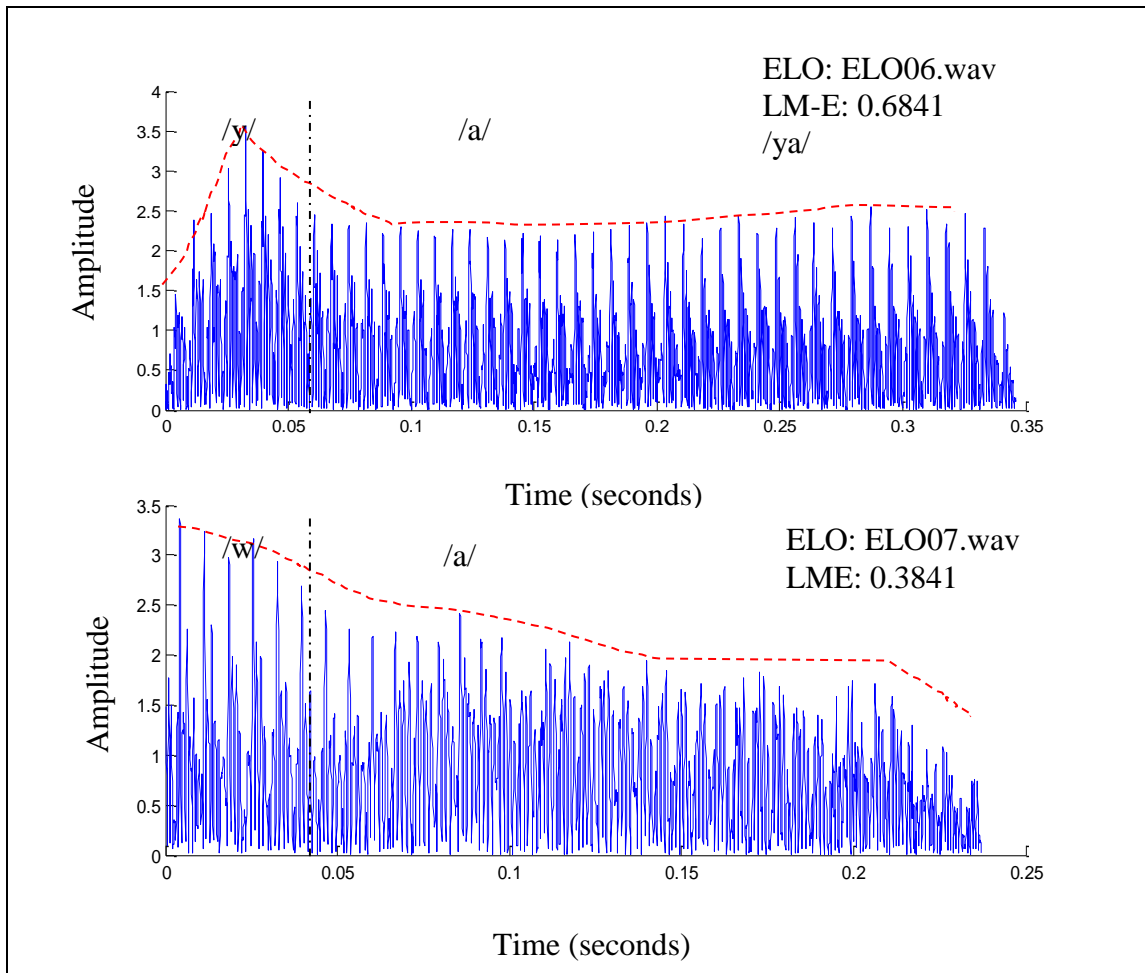


Figure 6.8: Example of misclassified elongation

In speech production, there is a transition between consonant to vowel causing the acoustic changes within the transition (Doellinger et al., 2011). According to Doellinger et al., (2011); Lisker & Abraham, (1967) the transition between consonant to vowel is due to the interval between the release burst and the onset of laryngeal pulsing. The transition from consonant to vowel in Malay language produced a unique phenomenon named as expressive intonation in this thesis. The graphical representation of the consonant to vowel transition is shown in Figure 6.9. Since there is no significant transition between consonant to vowel in the elongations depicted in Figure 6.8, a lower standard deviation of LM-E is derived. Thus, the standard deviation does not meet the acoustical rules of LM-E for elongation; they are misclassified as filled pause.

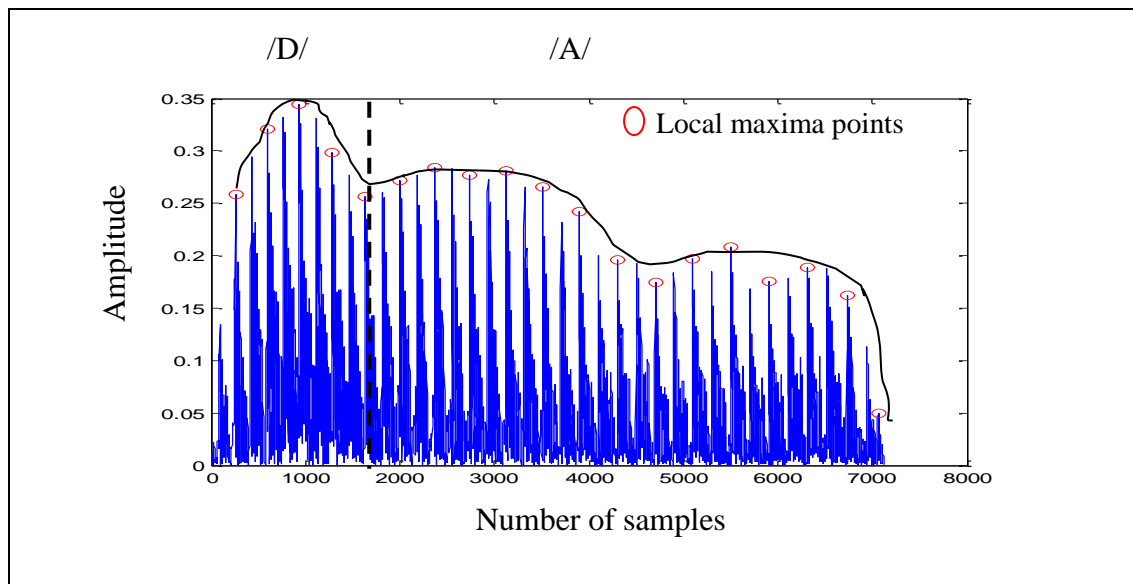


Figure 6.9: Consonant to vowel transition in elongation /da/

Some of the elongation starts with voiced consonant (i.e. /ga/, /da/, and /ni/) unvoiced consonant (i.e. /pi/, /tu/, and /ke/). There are also elongations uttered with semivowel (i.e. /ya/, /wa/). It is observable that there is no significant amplitude transition between consonant to vowel in many of the elongations of the 7th fold; thus causing lower LM-E standard deviation. The elongation that is in the form of semivowel is hardly to be correctly classified by LM-E. Most of the elongations cannot be correctly classified by using LM-E as the energy of the semivowel and the vowel of the filled pause do not differ significantly. Referring to Figure 6.8, there is no significant transition between /y/ to /a/ of the elongation /ya/ and /w/ to /a/ of the elongation /ya/. According to Wilson, (1986), the similar acoustical pattern between semivowel and vowel causing the detection of semivowels is a challenging task. In summary, several causes of misclassification done by LM-E are:

- i. A low volume of voice pronunciation by the speaker caused inaccurate representation of LM-E for filled pause.
- ii. Filled pause is uttered in an emotional state of mind such as angry, happy and doubt; producing expressive intonation in the filled pause utterances. Therefore, filled pause is misclassified as elongation as it possessed characteristic similar to elongation.
- iii. Insignificant transition between consonant to vowel in elongation; causing a low LME's standard deviation.

As stated earlier, the LM-E is associated with the speech energy (STE). Therefore, this research compares the performance of these two speech energy characteristics in differentiating filled pause and elongation. Since the filled pause is unvaried pronunciation of phonemes, the energy is constant. The consistency of the energy is measured based on STE's lower standard deviation (Veiga, 2011). In other words, the STE's standard deviation for filled pause is lower compared to elongation. The KDE of STE is shown in Figure 6.10. It is observable that the STE of filled pause and elongation overlaps from 0 to 210.

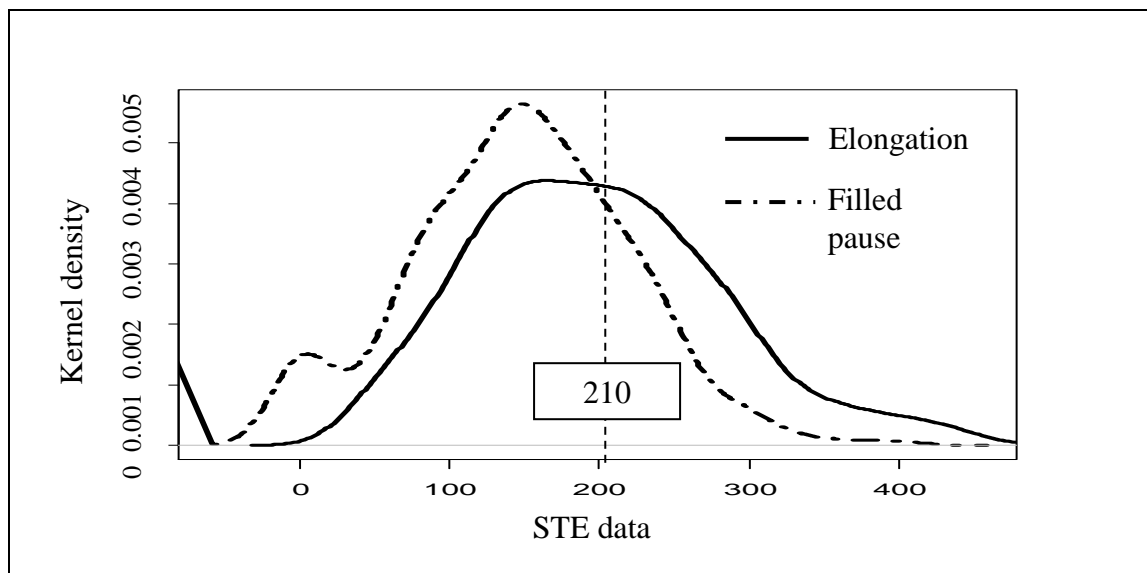


Figure 6.10: Kernel density estimation of STE for filled pause and elongation

The LM-E which is an exploitation of the speech energy, however managed to differentiate the elongation better compared to STE. The KDE of the LM-E is shown in Figure 5.5 (Chapter Five). It can be seen that there are less overlapping of LM-E value between filled pause and elongation. In this chapter, LM-E and ZCR have shown high accuracy averaging at 74% equivalently. Feature ranking in Chapter Five has also produced the same results of feature importance with ZCR and LM-E being the top-2 acoustical features. Therefore, these two acoustical features are chosen to be incorporated in the discriminative classification model.

6.5 DISCRIMINATIVE CLASSIFICATION PERFORMANCE EVALUATION

In the previous single feature classification, ZCR has scored the same accuracy of 74% as LM-E which is among the highest performance compared to the other acoustical features. Therefore, these two features are fused in the proposed discriminative classification model and the final posterior probability is computed using Naïve Bayes fusion classification (Refer Figure 5.10). Classification experiment is conducted using 3000 filled pause (i.e VAD_03) and 3000 elongation (i.e VAD_06) to determine its performance in classifying filled pause and elongation into two classes. The results of accuracy and F-measure of each fold performance are shown in Table 6.10 and Table 6.11.

The precision and recall that are used for F-measure calculation are shown in Table 6.12 and Table 6.13. The results are divided into three categories which are: Experiment 01, Experiment 02, and Experiment 03. In each experiment, the results of each fold of the cross validation and its average are presented.

The aim of Experiment01 is to test the performance of LM-E combination with the other acoustical feature. Results of Experiment01 suggest that the discriminative model (LM-E+ ZCR) denotes the highest average of 81% accuracy as seen in Table 6.10 compared to the other feature combination classification. The discriminative model shows the lowest performance of 75% and 76% in the third and fifth fold. In the previous classification of single feature for LM-E and ZCR, the lowest accuracy is 68% and 69%, respectively. However, with the combination of LM-E and ZCR, the average accuracy is increased by 13%. It can be seen that fold 01, fold 02, fold 06 had an above average performance by $\pm 3\%$, while fold 09 and fold 10 achieved increment of $\pm 1\%$. It is also observed that in Experiment01, the combination of LM-E and each acoustical feature significantly increased the single feature classification performance. For example, in the single feature performance, FF4 had the lowest average accuracy of 40%. When FF4 is fused with LM-E, a significant performance increment of 33% was observed, achieving an average of 73% accuracy. On average, each standard acoustical feature improved from as low as 10% for ZCR and as high as 83% for FF4 when the comparison is done between average single feature and multi-feature classification performance. The lowest multi-feature classification average of 73% is denoted in LM-E +FF4 and LM-E+MFCC. The recall rate, precision rate and F-

measure shown in Table 6.11 to 6.13 denote the same performance of LM-E+ZCR for experiment 01. The recall rate of LM-E+ZCR achieved the highest rate of 92% in the eighth fold for filled pause and 82% for elongation in the ninth fold. The highest precision rate of 88% for LM-E+ZCR of experiment 01 can be depicted in the sixth fold for filled pause and 90% for elongation in the eighth fold.

In Experiment 02, the purpose is to observe the performance of multi-feature classification without LM-E feature. Accuracy results are presented Table 6.10. It can be observed that the highest average accuracy of 77% was denoted in ZCR+STE, ZCR+FF2 and ZCR+FF1. However, ZCR+F0 scored the lowest accuracy of 73%. Although the combination of ZCR with other acoustical features has improved each of the single feature performance, it can be seen that the performance is significantly lower compared to the combination between LM-E and ZCR of 81%. The precision, recall and F-measure rate of Experiment 02 are shown in Table 6.11 to Table 6.13. From the tables, it is observable that Experiment 02 shows the same performance pattern as the accuracy for the multi-feature classification. The multi-feature classification performance is only able to achieve the highest recall rate of 81% shown in ZCR+STE depicted in the tenth fold of filled pause class. On the other hand, the highest precision of 80% achieved by the multi-feature classification of Experiment 02 is shown in ZCR+FF4 class and ZCR+MFCC for filled pause class. When the precision and recall is used in the F-measure calculation, the multi-feature classification only able to achieve the highest rate at 80% in the tenth fold of ZCR+STE. To prove that the discriminative model of LM-E+ZCR is effective for filled pause and elongation classification, all of the well-established acoustical features were also combined as shown in Experiment 03. In the previous filled pause classification researches, the STE, F0, MFCC, ZCR and FF are utilized as seen in Li et al., (2010), Li et al., (2010), Veiga et al., (2011). Therefore, these acoustical features are combined in this research to compare the classification performance. The result in Table 6.10 shows that combination of all features achieved a mere 71% average accuracy compared to 81% accuracy achieved by LM-E + ZCR. Even though all of the standard features are fully utilized, the performance is lower by an average of 10% compared to the discriminative model LM-E+ZCR performance, showing that more combination of acoustical features does not guarantee a higher classification performance. The correct combination of acoustical features is utmost critical in achieving higher accuracy rate.

Table 6.10:
10-folds CV accuracy (%) of discriminative classification using combined data

Experiment	Acoustical Features (combined)	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10	\bar{x} (Multi-feature)	\bar{x} (Single feature)	\bar{x} Increased performance of single-feature to multi-feature classification
Experiment01	LM-E+ZCR	83	83	75	81	76	83	80	80	82	82	81	ZCR: 74	10
	LM-E+STE	76	78	75	78	78	76	77	78	78	78	77	STE: 67	15
	LM-E+FF2	77	78	74	75	78	76	77	77	78	78	77	FF2: 68	13
	LM-E+FF1	75	74	73	75	75	74	75	77	75	74	75	FF1 : 65	15
	LM-E+F0	75	73	72	72	72	74	75	74	74	75	74	F0: 41	80
	LM-E+FF3	75	74	73	75	75	74	75	77	75	74	75	FF3: 62	21
	LM-E+MFCC	73	74	71	72	74	73	72	73	72	73	73	MFCC:53	38
	LM-E+FF4	71	74	71	72	73	73	72	73	73	73	73	FF4: 40	83
Experiment02	ZCR+FF1	77	78	75	76	77	77	77	76	76	78	77		
	ZCR+FF2	77	76	77	77	77	76	77	76	76	77	77		
	ZCR+FF3	75	74	73	74	74	75	75	76	75	74	75		
	ZCR+FF4	75	74	73	75	75	74	75	77	75	74	75		
	ZCR+F0	73	74	71	72	74	73	72	73	72	73	73		
	ZCR+MFCC	75	74	73	75	75	74	75	77	75	74	75		
	ZCR+STE	77	78	73	75	77	77	77	76	77	78	77		
Experiment03	ZCR+F0+FF+MFCC+ STE	72	70	70	71	71	72	72	70	71	71	71		

Table 6.11:
F-measure for 10 fold CV for combined data

Experiment	Acoustical Features	Fold1		Fold2		Fold3		Fold4		Fold5		Fold6		Fold7		Fold8		Fold9		Fold10	
		E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F
Experiment 01	LM-E+FF1	73	77	75	73	74	71	74	75	73	76	73	75	74	75	77	76	75	74	73	74
	LM-E+FF2	77	76	78	77	73	75	75	75	77	78	75	76	78	76	75	78	77	78	77	78
	LM-E+FF3	76	74	74	74	73	72	76	74	76	74	75	74	76	74	76	77	76	74	73	74
	LM-E+FF4	70	72	74	73	71	70	71	73	74	72	72	73	72	71	72	73	72	73	71	74
	LM-E+F0	75	73	71	74	72	72	70	74	71	73	74	73	73	76	74	73	74	73	74	76
	LME+MFCC	72	73	74	73	72	72	72	71	76	72	74	71	71	73	72	73	73	71	72	74
	LM-E+STE	75	77	77	78	72	75	78	77	77	78	77	75	75	78	77	78	77	78	77	78
	LM-E+ZCR	81	83	79	86	72	77	77	84	74	80	78	86	72	84	78	82	79	85	79	84
Experiment 02	ZCR+FF1	76	77	76	79	73	76	75	77	76	78	75	78	76	77	75	77	76	76	76	79
	ZCR+FF2	77	76	74	77	71	76	77	76	75	78	75	76	78	76	74	77	77	74	77	76
	ZCR+FF3	73	76	73	75	72	74	74	73	73	75	73	76	75	75	75	76	73	76	73	74
	ZCR+FF4	71	79	71	77	71	74	75	74	71	79	71	76	72	78	75	78	71	79	71	77
	ZCR+F0	70	76	72	75	70	71	70	74	72	75	70	76	70	74	70	76	70	74	70	76
	ZCR+MFCC	74	75	74	74	72	74	74	75	74	75	74	73	73	77	74	79	74	76	74	73
	ZCR+STE	76	77	78	77	72	74	76	74	76	78	76	77	76	78	75	77	76	78	76	80
Experiment 03	ZCR+F0+MFCC+FF+STE	71	73	70	69	69	70	72	69	70	72	72	71	72	71	70	70	70	72	72	70

E:ELO, F:FP

Table 6.12:

Precision of the 10-fold CV for combined data

Experiment	Acoustical Features	Fold1		Fold2		Fold3		Fold4		Fold5		Fold6		Fold7		Fold8		Fold9		Fold10	
		E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F
Experiment 01	LM-E+FF1	75	77	76	77	73	72	73	78	76	78	76	74	78	74	78	75	78	75	74	75
	LM-E+FF2	78	77	81	79	71	74	75	75	78	80	74	76	77	76	77	77	77	77	77	77
	LM-E+FF3	75	77	75	77	70	72	78	78	78	78	76	78	76	78	78	77	78	72	72	72
	LM-E+FF4	70	73	74	72	72	71	72	76	76	72	75	72	75	72	75	72	75	72	72	74
	LM-E+F0	75	77	70	74	72	73	72	76	72	74	72	74	72	74	72	72	72	72	72	78
	LME+MFCC	70	77	75	77	70	70	70	72	73	73	73	73	73	73	75	73	75	73	74	73
	LM-E+STE	73	77	79	76	71	76	79	81	75	81	75	75	75	77	77	78	77	78	77	78
	LM-E+ZCR	84	81	78	87	87	68	83	79	82	74	76	88	80	80	90	74	76	87	83	81
Experiment 02	ZCR+FF1	78	77	75	79	75	75	78	77	78	77	75	79	78	77	77	79	77	79	75	77
	ZCR+FF2	78	77	73	78	71	77	77	78	78	77	73	78	79	77	73	77	78	77	78	77
	ZCR+FF3	74	77	72	76	75	77	73	75	75	77	74	78	75	77	74	78	75	77	75	77
	ZCR+FF4	70	80	73	75	70	75	75	73	72	79	73	74	74	79	77	77	72	78	73	78
	ZCR+F0	70	77	73	78	70	70	70	75	73	77	70	77	70	76	70	77	71	77	70	77
	ZCR+MFCC	76	77	75	78	70	78	76	77	76	77	75	77	72	78	71	80	76	75	75	77
	ZCR+STE	75	77	79	77	73	76	75	77	77	80	73	77	75	77	73	77	75	77	75	79
Experiment 03	ZCR+F0+FF+MFCC+STE	70	72	71	68	70	69	73	66	69	73	70	74	73	69	71	72	69	74	69	69

E:ELO, F:FP

Table 6.13:

Recall rate of 10-fold CV for combined data

Experiment	Acoustical Features	Fold1		Fold2		Fold3		Fold4		Fold5		Fold6		Fold7		Fold8		Fold9		Fold10	
		E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F	E	F
Experiment 01	LM-E+FF1	71	78	75	70	75	71	75	73	70	75	70	76	70	76	76	78	73	74	73	74
	LM-E+FF2	76	75	75	75	75	76	75	75	76	77	76	77	79	77	74	79	77	79	77	79
	LM-E+FF3	78	72	73	72	76	73	75	70	75	70	75	70	77	70	75	77	75	77	75	77
	LM-E+FF4	71	72	75	75	71	70	71	70	72	73	70	74	70	70	70	75	70	75	70	75
	LM-E+F0	71	78	72	75	72	72	68	72	70	72	76	73	75	78	76	75	76	75	76	75
	LME+MFCC	74	70	74	70	74	74	74	71	79	71	76	70	70	74	70	74	72	70	70	76
	LM-E+STE	77	78	75	80	74	74	77	74	79	76	79	76	76	79	78	78	78	78	78	78
	LM-E+ZCR	79	86	80	86	61	90	71	89	67	86	80	85	66	89	69	92	82	83	76	87
Experiment 02	ZCR+FF1	74	78	77	80	71	78	72	78	74	79	76	78	74	78	73	75	76	73	77	80
	ZCR+FF2	77	75	76	76	72	75	78	75	73	79	77	75	77	75	76	78	76	72	76	75
	ZCR+FF3	73	75	75	75	70	72	75	72	71	74	73	74	76	73	77	74	72	75	72	72
	ZCR+FF4	73	78	70	79	73	73	76	75	71	80	70	78	71	77	74	79	70	80	70	76
	ZCR+F0	71	75	71	73	71	72	71	73	71	74	71	75	71	73	71	75	70	72	71	76
	ZCR+MFCC	73	73	74	70	74	70	72	73	73	73	73	70	74	76	77	78	73	77	74	70
	ZCR+STE	77	78	77	78	72	73	77	72	75	76	79	78	77	79	77	78	77	79	77	81
Experiment 03	ZCR+F0+FF+MFCC+STE	72	75	70	70	68	72	71	73	72	72	74	69	70	73	70	70	72	70	76	72

E:ELO, F:FP

The proposed acoustical feature LM-E is also deemed as a contributing factor in the discriminative model. As shown in Experiment 02 and 03, the absence of LM-E in the multi-feature classification caused a drop of accuracy rate to 77% (i.e. ZCR+STE) and 71% (ZCR+F0+FF+MFCC+STE). On the other hand, LM-E+ZCR managed to achieve 81% accuracy rate in Experiment 01. The results of Experiment 03 for recall, precision and F-measure are shown in Table 6.11 to Table 6.13.

It is clearly observed that the combination of more features cannot increased the recall, precision and F-measure. The highest recall rate of 76% is achieved in the tenth fold of the Experiment 03 for filled pause class, while the highest precision of 74% can be seen in the sixth fold for filled pause class. The harmonic mean of recall and precision which is the F-measure indicated the highest rate of 71% in the first fold for filled pause class.

To show the mechanism of the feature based Naïve Bayes combination of the discriminative model, two samples of data is chosen; ELO07.wav and FP11.wav. These examples are the same samples illustrated in *Section 5.2*. The conditional probability densities were estimated from the Kernel density estimation as shown in Chapter Five for LM-E and ZCR. The posterior probability is calculated using equation (5.2). The information of the data is shown in Table 6.14.

Table 6.14:
Information of probability density estimation of FP11.wav and ELO07.wav

Information	FP11.wav	ELO07.wav
LM-E value of speech segment	0.93	0.38
ZCR value of speech segment	22.0	14.2
LM-E conditional probability of FP	0.51	0.40
LM-E conditional probability of ELO	1.50	0.20
ZCR conditional probability of ELO	0.04	0.90
ZCR conditional probability of FP	0.14	0.03
Posterior probability of FP	0.54	0.06
Posterior probability of ELO	0.46	0.94

From the above example, it can be seen that the combination between LM-E and ZCR complements each other and successfully classifies the disfluencies FP11.wav and ELO07.wav. In the FP11.wav example, the LM-E's standard deviation value is in the elongation class (refer to estimated acoustical parameter Table 5.1). Previously, these examples are misclassified to the wrong class by using LM-E single feature classification. However, when the ZCR value is combined with the LM-E, the filled pause is correctly classified in its own class. The results further suggest that, the

combination between LM-E and ZCR increase the probability of correct classification for filled pause and elongation.

6.6 CLASSIFICATION PERFORMANCE COMPARISON WITH ENGLISH LANGUAGE DATASETS

The discriminative classification model of ZCR and LM-E is further tested to test its robustness on a benchmarked English language filled pause and elongations. The data is gathered from Linguistic Data Consortium (LDC). Due to data availability constrain, only 220 filled pauses and 220 elongations are used. The types of English language elongations are shown as in Table 6.15.

Table 6.15:
English elongations

English Word	Elongated	Structure	English Word	Elongated	Structure
There		CCVCV	So		CV
Last		CVCC	Trail		CCVVC
Over		VCVC	About		VCVVC
That		CCVC	Showers		CCVCVCC
All		VCC	Now		CVC
They		CCVC	Aloud		VCVVC
Call		CVCC	Also		VCCV
Five		CVCV	And		VCC
Burn		CVCC	Anyway		VCCCVCC
Usually		VCVVCCC	Today		CVCVC

C: Consonant; V: Vowel

It can be seen from Table 6.15 that the structure of elongations are different with Malay language. Some of the elongations are in the form of words and some in syllables. For example, the words such as *there*, *last*, *and*, *so*, and *all* are elongated as a whole while the words *over*, *aloud*, *also*, *anyway* and *today* are elongated at the last syllables. English elongation can also be found at the first syllable /*sho*/ of the word such as the word *showers*.

Figure 6.11 shows an elongated whole word, ‘that’, spoken by a male speaker. It can be seen that the elongated word is uttered at a duration of more than 300ms. As has been stated earlier, filled pause and elongation are normally uttered at longer duration (i.e. >200ms).

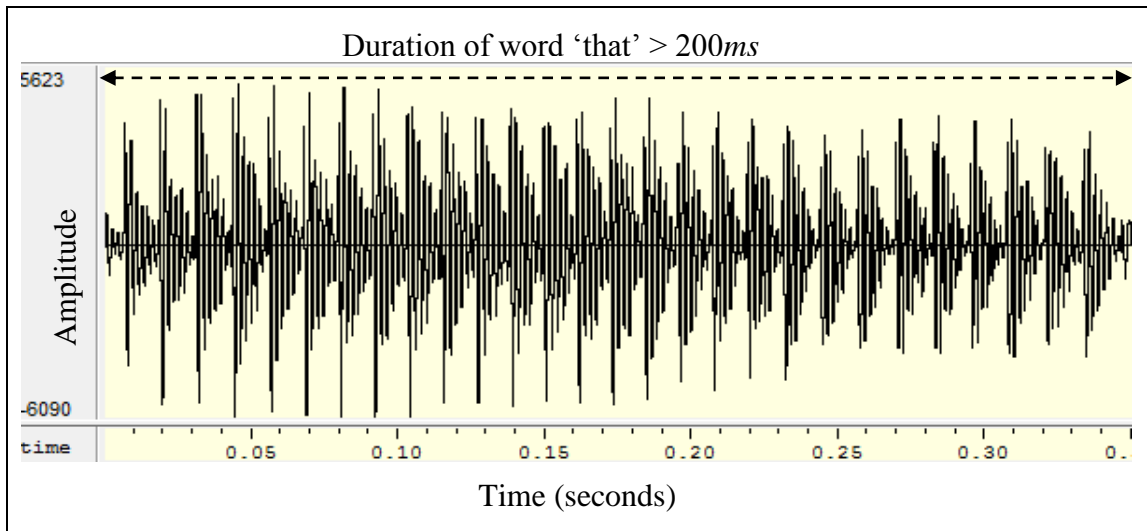


Figure 6.11: English language elongation of the word 'that'

Figure 6.12 shows an elongation that appears at the first and second syllable of the word *showers*. It is observable that both the first and second syllables /sho/ and /wers/ are uttered at more than 200ms making the overall duration of the word ~500ms. In Figure 6.13, elongation occurs at the last syllable /wer/. The last syllable is elongated at 290ms and stop 600ms making the overall duration for the elongation is 310ms.

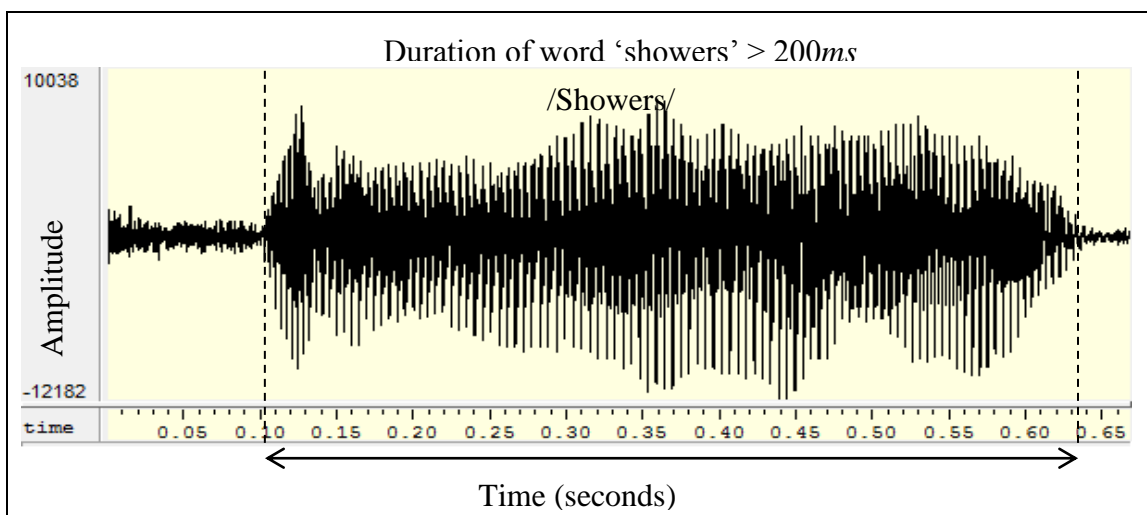


Figure 6.12: English language elongation of the word 'showers'

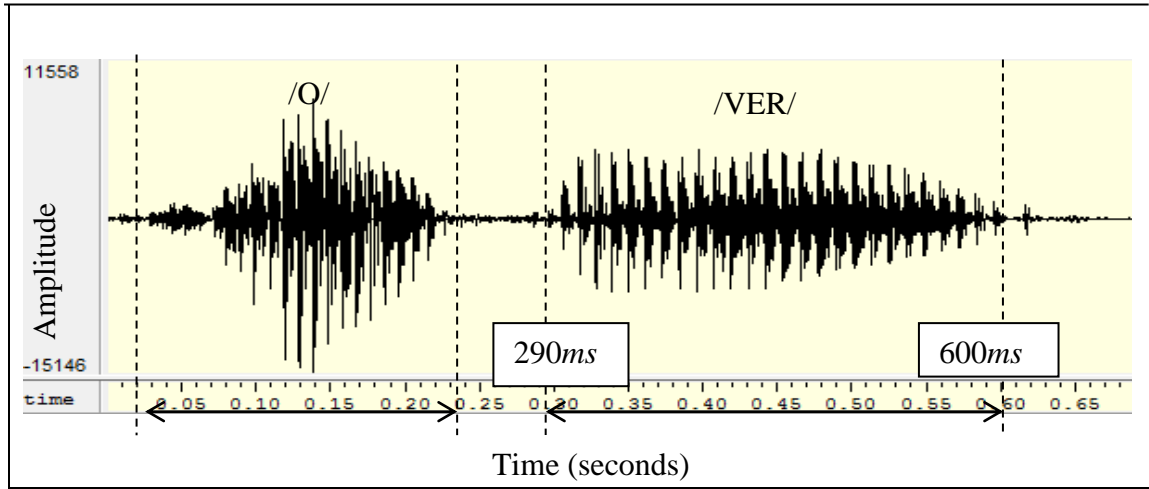


Figure 6.13: English language elongation of the syllable 'ver'

The English datasets is evaluated using 10-folds cross validation. The results are shown in Table 6.16.

Table 6.16:

Discriminative model (LM-E+ZCR) classification performance on English data

Fold	F-Measure %		Accuracy %
	ELO	FP	
1	79	81	80
2	74	76	75
3	70	70	70
4	62	75	69
5	71	78	75
6	76	74	75
7	75	57	66
8	70	74	72
9	75	60	68
10	78	60	69
Average	73	71	72

From Table 6.15, the highest F-measure of 79% of the discriminative model (LM-E+ZCR) for elongation and 81% for filled pause class can be seen in the first fold. The highest accuracy of 80% is also denoted in the first fold. The discriminative model has able to achieve overall performance at average of 73% and 71 % F-measure for elongation and filled pause class respectively and 72% accuracy.

Table 6.17:
Comparison of discriminative model (LM-E+ZCR) classification on Malay and English language datasets

Dataset	Discriminative model	Average F-measure%		Average accuracy%
		ELO	FP	
Malay	ZCR_LM-E	77	83	81
English	ZCR_LM-E	73	71	72

Table 6.17 shows the average accuracy of English language dataset is lower by 9% compared to Malay dataset. From the observation, English language filled pause is easier to be classified compared to elongations. Elongated words such as /there/ and /they/ are among the hardest to be classified using the discriminative model of ZCR_LM-E. The unvoiced th in /there/ and /they/ are pronounced with unvibrated vocal cord thus, treated as unvoiced sound. Compared to the Malay language datasets, the unvoiced type of elongated syllable is unable to be correctly classified by the discriminative model. The filled pause acoustical features are already well-established for languages such as English, Mandarin and Portuguese. However, more exploration is needed for elongation. Eklund, (2001) has stated that more studies need to be done cross-linguistically to get the insight of humans' elongations as they are language specific.

6.7 SUMMARY

In summary, ZCR and LM-E acoustical features are shown to perform better than other acoustical features in the 10-fold cross-validation single-features classification. LM-E feature is also proven to discriminate elongation from filled pause. Based on these findings, a discriminative classification model using Naïve Bayes is constructed. The discriminative classification model combining ZCR and LM-E features successfully classified filled pause and elongation with better performance compared to single feature and combination of all standard acoustical features. The combination of ZCR and LM-E complements each other in the filled pause and elongation classification, thus increasing the number of correct classification. It is also summarized that the relevant features that need to be fused are important compared to the number of features fused in each classification.

CHAPTER SEVEN

CONCLUSION AND FUTURE WORKS

7.1 INTRODUCTION

This chapter summarized the research presented in this thesis and discusses the direction of future work. This thesis has addressed the challenge of classifying the Malay language filled pause and elongation into two classes. An improved local maxima of speech energy (LM-E) feature is introduced to describe the expressive intonation of Malay language elongation. When fused with zero crossing rate (ZCR), and composed in Naïve Bayes discriminative model, an average accuracy of 81% is achieved for classification of filled pause and elongation. This thesis also introduced a new set of acoustical features-based rules incorporated in Kernel density estimation of the proposed model. This chapter contains a summary of the work executed in this research, the main contribution and achievements, its significance and the limitations of the proposed framework. The future work is discussed at the end of this chapter.

7.2 REVIEW OF OBJECTIVES

The problems associated with the filled pause and elongation classification are emphasized in Chapter 1, Chapter 2 and Chapter 3. The three main objectives which formulated for this research are:

Objective 1: To create Malay language’s filled pause and elongation dataset from spontaneous speech.

This objective is achieved through the construction of Malay language filled pauses and elongations datasets. Each of the filled pause and elongation went through well-established voice activity detection method based on the threshold of energy and higher order differences of the speeches. **(The discussion is in Chapter Three).**

Objective 2: To produce a new energy feature for filled and elongation.

This objective is achieved through comparing a well-established energy extraction technique (STE) and local-maxima extraction. Three types of local maxima extraction

techniques are applied and compared to produce new local maxima of the speech energy. **(The discussion of this objective and its solution are elaborated in Chapter Four).**

Objective 3: To model the discriminative properties of filled pause and elongation acoustical features.

The objective is fulfilled through the probability estimation of the acoustical features density by using Kernel density method. The Kernel-probability density estimation provides the ranges of probability density of the ZCR and LM-E that is used in the Naïve Bayes classification. **(The discussion is in Chapter Five)**

7.3 SUMMARY OF RESEARCH FINDINGS

Filled pause detection and classification is not a new research area which has attracted attention since 1994. Shriberg (1994) investigated filled pause as one of the disfluencies in English language spontaneous speech by proving that filled pause as the most occurred disfluencies in spontaneous speech. Despite its linguistic role in helping the speakers in maintaining the conversation, the high occurrences of filled pause in spontaneous speech also degrades automatic speech recognition's (ASR) performance. Li et al., (2008) stated that the main factor of filled pause misclassification is associated with elongation's existence. Elongation mimics the filled pause since their acoustical features patterns are similar. Therefore, many filled pause researcher grouped filled pause and elongation into the same class. However, by treating elongation as filled pause can change the semantic meaning of a sentence since elongation has its own vocabulary meaning while filled pause does not. The need to classify them separately is also mutually agreed by Li et al., (2008); Li et al., (2010); Veiga et al., (2011); Verkhodanova & Shapranov (2014). One of the most important factors for better filled pause and elongation classification is the selection of the most appropriate acoustical features that is able to discriminate the filled pause and elongation.

Preliminary investigation to rank the acoustical features is conducted measuring the highest mean Z-score using Random Forest algorithm. Results suggested that ZCR and LM-E are the two most contributing features for Malay language filled pause and elongation. In Chapter Four the introduction of local

maxima of the speech energy (LM-E) is shown to contribute to a better classification performance by producing higher accuracy. Other than LM-E, ZCR is also another contributing feature. The use of LM-E at certain threshold settings enabled it to represent the expressive intonation characteristic and eventually discriminates the elongations among filled pauses. On the other hand, ZCR has the ability in discriminating elongation from filled pause compared to other features, because of its capability in detecting vowel and consonant. A new set of acoustical rules are developed in Chapter Five and incorporated in the Bayes theorem classification for single feature classification.

Overall, nine acoustical features are extracted and evaluated. The ZCR and LM-E are among the most contributing acoustical features by scoring 43 and 42 mean Z-score respectively. The second phase evaluation by single Bayes classification also proves that ZCR and LM-E are the best acoustical features by scoring 26% error rates. The next phase is then to fuse the most contributing features of LM-E and ZCR to build the filled pause and elongation discriminative model by using product rule of Naïve Bayes based on the Kernel density estimation. The discriminative model improves the classification performances compared to single pattern classification.

Chapter Five also presented the discriminative model of ZCR and LM-E that is built using product rule of Naïve Bayes based on Kernel density estimation. Three experiments testing fusion of acoustical features are conducted using the proposed discriminative model. Findings showed that fusion of LM-E and ZCR produced the highest accuracy compared to fusion of other acoustical features.

7.4 SIGNIFICANT CONTRIBUTION OF THE RESEARCH

This research introduced a discriminative model to classify filled pause and elongation into its own group based on rigorous acoustical feature selection and Kernel based Naïve Bayes method. Having achieved the objectives of this research, several key contributions have been made to the filled pause and elongation discriminative classification. The contributions of this research are summarized as follows:

- i. A collection of Malay language filled pauses and elongations are created. Malay language is still considered as an under resourced language (Besacier et al., 2013) due to its lack of available digital resources. This data collection is carefully chosen from the Malaysian Parliamentary Hansard document (MPHD) of the year 2008 and pre-processed to produce exact representation of the disfluencies. These datasets can be used for further research of automated speech recognition and add to the digital resources of Malay language.
- ii. An improved local maxima of the speech energy (LM-E) as another acoustical feature is proposed for better filled pause and elongation classification. Each of the acoustical features is evaluated based on Boruta-Random Forest feature selection method and single feature-based classification.
- iii. The introduction of LM-E has increased the accuracy of the single feature-based classification and significantly improved the classification performance when fused with ZCR together using Kernel based Naïve Bayes method. LM-E is further improved using a new threshold calculation technique.
- iv. A new set of acoustical rules are identified and incorporated in the Kernel density estimation function to generate the conditional probability of each feature.
- v. A Naïve Bayes discriminative classification model is constructed by fusing LM-E and ZCR. The model is able to achieve an average 81% accuracy rate for classification of filled pause and elongation

7.5 RESEARCH LIMITATION

When using the proposed discriminative classification model, the elongation and filled pause need to be more than $200ms$ for the classifier to classify. However, when filled pause with duration of lower than $200ms$ is used, another problem arise which is the confusion with short word. Although many of the previous works has benchmarked the minimum filled pause duration is $200ms$, the existence of filled pause with shorter duration still occurs and need to be taken care of.

There are many types of disfluencies in spontaneous speech. However, this thesis only focus on three types of filled pause which are ‘aaa’, ‘eee’ and ‘eem’ that are identified as the most occurred filled pause types in the MPHD data. Therefore, the discriminative model is developed to classify these types of filled pause and elongations.

The proposed discriminative model is also unable to correctly classify expressive intonation filled pause. The speech that were spoken by multi-racial speakers are also one of the factor that affect the expressive intonation and slang of the filled pause.

7.6 FUTURE RESEARCH ENHANCEMENT

Although this research presented the potential of the proposed discriminative model in classifying filled pause and elongation, further research remains to be done. In this research, the discriminative model is built from the Kernel-based Naïve Bayes method by utilizing the product rules in order to fuse the features. This fusion technique is categorized as decision fusion (Xu et al., 1992). Instead of decision fusion, there are many other types of fusion available in the literature such as feature fusion, data fusion and multilevel fusion (Castanedo, 2013). These other fusion types may be used in the future research to compare the most effective method in combining filled pause and elongation features.

The use of local maxima of the speech energy (LM-E) is easily affected by the threshold setting. In Chapter Five, various threshold techniques for LM-E were tested. Through observation, the midrange that were used in the threshold setting provides better solution for LM-E extraction as the classification performance for both filled pause and elongation became better compared to the standard threshold. Although the midrange threshold setting is better, the extraction of LM-E still can be improved. For better LM-E extraction, a tedious trial an error is needed for each specific filled pause and elongation. An adaptive threshold that can fit with each of the speech samples is needed to overcome this limitation.

Finally, the classification of filled pause and elongation can be extended into the detection process which involves filled pause detection among normal words and other spontaneous disfluencies such as repetition and sentence restart. Then, the detection process can be embedded into the automatic speech recognition system to

see the impact towards the ASR's performance. Previously, Kaushik, (2010) has proven that the removal of filled pause and elongation reduced the error rate of the ASR. Therefore, this is a good challenge that requires more works to be done to enable the future research accomplishments.

REFERENCES

- Abbas, E. I., & Refeis, A. A. (2013). Influence of Noisy Environment on the Speech Recognition Rate Based on the Altera FPGA.
- Aibinu, A. M., Salami, M. J. E., Najeeb, A. R., Azeez, J. F., & Rajin, S. A. K. (2011). Evaluating the effect of voice activity detection in isolated Yoruba word recognition system. In *Mechatronics (ICOM), 2011 4th International Conference On* (pp. 1-5). IEEE.
- Al-Alaoui, M. A., Al-Kanj, L., Azar, J., & Yaacoub, E. (2008). Speech recognition using artificial neural networks and hidden Markov models. *IEEE Technology and Engineering Education (ITEE)*, 3(3), 77-86.
- Al-Haddad, S. A. R., Ishak, K. A., Samad, S. A., Abid, A. O., & Noor, A. H. (2008). Robust digit recognition with dynamic time warping and recursive least squares. In *Information Technology, 2008. ITSIM 2008. International Symposium on* (Vol. 2, pp. 1-8). IEEE.
- Altınçay, H. (2005). On Naive Bayesian fusion of dependent classifiers. *Pattern Recognition Letters*, 26(15), 2463-2473.
- Altwaijry, H. (2013). Bayesian based intrusion detection system. In *IAENG Transactions on Engineering Technologies* (pp. 29-44). Springer Netherlands.
- Atal, B. S., & Rabiner, L. R. (1976). A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(3), 201-212.
- Audhkhasi, K., Kandhway, K., Deshmukh, O. D., & Verma, A. (2009). Formant-based technique for automatic filled-pause detection in spontaneous spoken English. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (pp. 4857-4860). IEEE.
- Avilés-Arriaga, H., Sucar-Succar, L., Mendoza-Durán, C., & Pineda-Cortés, L. (2011). A comparison of dynamic Naive bayesian classifiers and hidden markov models for gesture recognition. *Journal of applied research and technology*, 9(1), 81-102.

- Bachu, R. G., Kopparthi, S., Adapa, B., & Barkana, B. D. (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) Zone Conference Proceedings* (pp. 1-7).
- Bartkova, K. (2005). Prosodic cues of spontaneous speech in French. In *Disfluency in Spontaneous Speech*.
- Begum, M., Ailon, R. N., Zainuddin, R., Don, Z. M., & Knowles, G. (2008). Prosody generation by integrating rule and template-based approaches for emotional Malay speech synthesis. In *TENCON 2008-2008 IEEE Region 10 Conference* (pp. 1-6). IEEE.
- Bertot, E. M., Beaujean, P. P., & Vendittis, D. (2014). Refining Envelope Analysis Methods using Wavelet De-Noising to Identify Bearing Faults. *Uropean Conference Of The Prognostics And Health Management Society* (pp 1–8).
- Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56, 85-100.
- Bouckaert, R. R. (2005). Naive Bayes classifiers that perform well with continuous variables. In *AI 2004: Advances in Artificial Intelligence* (pp. 1089-1094). Springer Berlin Heidelberg.
- Bourlard, H., & Morgan, N. (1998). Hybrid HMM/ANN systems for speech recognition: Overview and new research directions. In *Adaptive Processing of Sequences and Data Structures* (pp. 389-417). Springer Berlin Heidelberg.
- Bowman, A. (1984). An alternative method of cross-validation for the smoothing of kernel density estimates. *Biometrika* 71, 353–360.
- Buller, D. B., & Burgoon, J. K. (1996). Interpersonal Deception Theory. *Communication Theory*, 6(3), 201-242.
- Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajič, J., Oard, D., & Zhu, W. J. (2004). Automatic recognition of spontaneous speech for access to multilingual oral history archives. *Speech and Audio Processing, IEEE Transactions on*, 12(4), 420-435.
- Cai, R., Lu, L., Zhang, H. J., & Cai, L. H. (2003). Highlight sound effects detection in audio stream. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on* (Vol. 3, pp. III-37). IEEE.

- Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, 2013.
- Cheng, J., Xiong, W., Gu, Y., Chia, S. C., Wang, Y., & Lim, J. H. (2015). A Three-Color Coupled Level-Set Algorithm for Simultaneous Multiple Cell Segmentation and Tracking. In *Computer Vision--ACCV 2014* (pp. 268-283). Springer International Publishing.
- Chetouani, M., Faundez-Zanuy, M., Gas, B., & Zarader, J. L. (2009). Investigation on LP-residual representations for speaker identification. *Pattern Recognition*, 42(3), 487-494.
- Chong, T. Y., Xiao, X., Tan, T. P., Chng, E. S., & Li, H. (2012). Collection and annotation of Malay conversational speech corpus. In *Speech Database and Assessments (Oriental COCOSDA), 2012 International Conference on* (pp. 30-35). IEEE.
- Clark, H. H., & Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73-111.
- Collobert, R., Bengio, S., & Bengio, Y. (2002). A parallel mixture of SVMs for very large scale problems. *Neural computation*, 14(5), 1105-1114.
- Costa, E., Lorena, A., Carvalho, A. C. P. L. F., & Freitas, A. (2007). A review of performance evaluation measures for hierarchical classifiers. In *Evaluation Methods for Machine Learning II: papers from the AAAI-2007 Workshop* (pp. 1-6).
- Cowell, F. A. and E. Flachaire. (2015). Statistical methods for distributional analysis. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution, Volume 2A, Chapter 6*. New York: Elsevier Science B. V.
- Cremer, F., Schutte, K., Schavemaker, J. G., & den Breejen, E. (2001). A comparison of decision-level sensor-fusion methods for anti-personnel landmine detection. *Information fusion*, 2(3), 187-208.
- Cucu, H. (2011). Towards a speaker-independent, large-vocabulary continuous speech recognition system for Romanian. *Teză de doctorat, Universitatea Politehnica din București, România*.
- Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1952). Formant Transitions and Loci as Acoustic Correlations of Place of Articulation in American Fricatives.

- Deme, A., & Markó, A. (2013). Lengthenings and Filled Pauses in Hungarian Adults' and Children's Speech. In *Sixth Workshop on Disfluency in Spontaneous Speech*.
- Deng, L., & O'Shaughnessy, D. (2003). *Speech processing: a dynamic and optimization-oriented approach*. CRC Press.
- Doellinger, M., Burger, M., Hoppe, U., Bosco, E., & Eysholdt, U. (2011). Effects of consonant-vowel transitions in speech stimuli on cortical auditory evoked potentials in adults. *The open neurology journal*, 5, 37.
- Domingos, P., Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learning* 29, 103–130.
- Donthu, N. (1991). Comparing market areas using kernel density estimation. *Journal of the Academy of Marketing Science*, 19(4), 323-332.
- Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning: proceedings of the twelfth international conference* (Vol. 12, pp. 194-202).
- Eklund, R. (2001). Prolongations: A dark horse in the disfluency stable. In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*.
- Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, 46(9), 913-920.
- Elkan, C. (2012). Evaluating Classifiers.
- El-Imam, Y. A., & Don, Z. M. (2000). Text-to-speech conversion of standard Malay. *International Journal of Speech Technology*, 3(2), 129-146.
- Enxhi, S. Y., Hoon, T. B., & Fung, Y. M. (2012). Speech disfluencies and mispronunciations in English oral communication among Malaysian undergraduates. *International Journal of Applied Linguistics and English Literature*, 1(7), 19-32.
- Fogerty, D., & Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, 131(2), 1490-1501.
- Fook, C. Y., Hariharan, M., Yaacob, S., & Adom, A. (2012). A review: Malay speech recognition and audio visual speech recognition. In *Biomedical Engineering (ICoBE), 2012 International Conference on* (pp. 479-484). IEEE.

- Furui, S., Nakamura, M., Ichiba, T., & Iwano, K. (2005). Analysis and recognition of spontaneous speech using Corpus of Spontaneous Japanese. *Speech Communication, 47*(1), 208-219.
- Gabrea, M., & O'Shaughnessy, D. (2000). Detection of filled pauses in spontaneous conversational speech. In *Sixth International Conference on Spoken Language Processing*.
- Ganapathy, S. (2012). *Signal analysis using autoregressive models of amplitude modulation* (Doctoral dissertation, Johns Hopkins University).
- Garg, D., Kaur, S., & Arora, D. (2012). Comparative Analysis of Speech Processing Techniques for Gender Recognition. *International Journal of Advances in Electrical and Electronics Engineering (IJAEEE, ISSN: 2319-1112), 1*(02), 278-283.
- Garg, G., & Ward, N. (2006). *Detecting filled pauses in tutorial dialogs*. (Technical Report, The University of Texas at El Paso).
- Goto, M., Itou, K., & Hayamizu, S. (1999). A real-time filled pause detection system for spontaneous speech recognition. In *Eurospeech*.
- Guidoum, A. C. (2015). Kernel Estimator and Bandwidth Selection for Density and its Derivatives.
- Gunawan, T. S., Abushariah, A. A., & Khalifa, O. O. (2011). English digits speech recognition system based on hidden Markov Models.
- Hall, M. (2007). A decision tree-based attribute weighting filter for Naive Bayes. *Knowledge-Based Systems, 20*(2), 120-126.
- Han, W., Hon, K. W., Chan, C. F., Lee, T., Choy, C. S., Pun, K. P., & Ching, P. C. (2003). An HMM-based speech recognition IC. In *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on* (Vol. 2, pp. II-744). IEEE.
- Hansen, B. E. (2009). Lecture notes on nonparametrics. *Lecture notes*.
- Hariharan, M., Chee, L. S., Ai, O. C., & Yaacob, S. (2012). Classification of speech dysfluencies using LPC based parameterization techniques. *Journal of medical systems, 36*(3), 1821-1830.
- Heeman, P. A., & Allen, J. F. (1999). Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics, 25*(4), 527-571.

- Holmes, J. & Holmes, W. (2001), *Speech Synthesis and Recognition*, 2th ed., Taylor & Francis, London.
- Honal, M., & Schultz, T. (2005). Automatic Disfluency Removal on Recognized Spontaneous Speech-Rapid Adaptation to Speaker Dependent Disfluencies. In *ICASSP (1)* (pp. 969-972).
- Hu, Y. (2009). *Detecting non-speech in dysarthric speech* (Doctoral dissertation, Department of Computer Science, University of Sheffield).
- Humpherys, S. L. (2010). A system of deception and fraud detection using reliable linguistic cues including hedging, disfluencies, and repeated phrases (Doctoral dissertation, Department of Business Administration, The University of Arizona).
- Hussain, A. (1997). The development of a phoneme based Malay speech recognition system using the modular artificial neural network approach.
- Iqbal, M. (2014). Network intrusion detection with Naïve Bayes Classification and Self Organizing Maps (Master Thesis, University Of Technology Sydney).
- Ito, M. R., & Donaldson, R. W. (1971). Zero-crossing measurements for analysis and recognition of speech sounds. *Audio and Electroacoustics, IEEE Transactions on*, 19(3), 235-242.
- Izzad, M., Jamil, N., & Bakar, Z. A. (2013). Speech/non-speech detection in Malay language spontaneous speech. In *Computing, Management and Telecommunications (ComManTel), 2013 International Conference on* (pp. 219-224). IEEE.
- Jain, J. R. (1986). Amplitude normalization and its application to speech coding. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86*. (Vol. 11, pp. 833-836). IEEE.
- Jalil, M., Butt, F. A., & Malik, A. (2013). Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. In *Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2013 International Conference on* (pp. 208-212). IEEE.
- McHugh, J. (2000). Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. *ACM transactions on Information and system Security*,3(4), 262-294.

- Kammoun, M. A., Gargouri, D., Frikha, M., & Hamida, A. B. (2006). Cepstrum vs. LPC: A Comparative Study for Speech Formant Frequencies Estimation. *GESTS Int. Trans. Commun. Signal Proc*, 9(1), 87-102.
- Karpiński, M. (2013). Acoustic Features of Filled Pauses in Polish Task-Oriented Dialogues. *Archives of Acoustics*, 38(1), 63-73.
- Kaushik, M., Trinkle, M., & Hashemi-Sakhtsari, A. (2010). Automatic detection and removal of disfluencies from spontaneous speech. In *Proceedings of the 13-th Australasian International Conference on Speech Science and Technology (SST). Melbourne, Australia* (pp. 98-101).
- Keerio, A., Mitra, B. K., Birch, P., Young, R., & Chatwin, C. (2009). On preprocessing of speech signals. *International Journal of Signal Processing*, 5(3), 216-222.
- Khan, M. Farhan, A. Ali. (2011). Speech Recognition – Increasing Efficiency of Support Vector Machines. *International Journal of Computer Applications*, Vol. 35, No. 7, 2011, pp. 17-21.
- Khojastehrad, S. (2012). Gender Differentiation in the Application of Hesitation Strategies among EFL Learners. *Advances in Asian Social Science*, 1(2), 205-211.
- Kim, C., Seo, K. D., & Sung, W. (2006). A robust formant extraction algorithm combining spectral peak picking and root polishing. *EURASIP Journal on Applied Signal Processing* (33-33).
- Kitayama, K., Goto, M., Itou, K., & Kobayashi, T. (2003). Speech starter: noise-robust endpoint detection by using filled pauses. In *INTERSPEECH*.
- Kitagawa, A., Watanabe, A., & Kumaki, H. (2011). A Path to developing oral proficiency : Speaking rate , silent pauses and fillers. *Pan-Pacific Association of Applied Linguistics* (261-268).
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, 36(i11).
- Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. *BMC bioinformatics*, 15(1), 8.
- Krüger, S. E., Schafföner, M., Katz, M., Andelic, E., & Wendemuth, A. (2006). Mixture of support vector machines for hmm based speech recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on* (Vol. 4, pp. 326-329). IEEE.

- Lee, L. W., Low, H. M., & Mohamed, A. R. (2013). A Comparative Analysis of Word Structures in Malay and English Children's Stories. *Pertanika Journal of Social Sciences & Humanities*, 21(1).
- Lee, J. S., & Park, C. H. (2008). *Adaptive decision fusion for audio-visual speech recognition*. INTECH Open Access Publisher.
- Lee, T. L., He, Y. F., Huang, Y. J., Tseng, S. C., & Eklund, R. (2004). Prolongation in spontaneous Mandarin. In *INTERSPEECH*.
- Li, X., & Stern, R. M. (2003). Feature generation based on maximum classification probability for improved speech recognition. In *INTERSPEECH*.
- Li, K., Swamy, M. N. S., & Ahmad, M. O. (2005). An improved voice activity detection using higher order statistics. *Speech and Audio Processing, IEEE Transactions on*, 13(5), 965-974.
- Li, Y. X., He, Q. H., & Li, T. (2008). A novel detection method of filled pause in mandarin spontaneous speech. In *Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on* (pp. 217-222). IEEE.
- Li, Y. X., He, Q. H., Li, W., & Wang, Z. F. (2010). Two-level approach for detecting non-lexical audio events in spontaneous speech. In *Audio Language and Image Processing (ICALIP), 2010 International Conference on* (pp. 771-777). IEEE.
- Lin, C. K., & Lee, L. S. (2009). Improved features and models for detecting edit disfluencies in transcribing spontaneous mandarin speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(7), 1263-1278.
- Lisker, L. & Abraham, A. S. (1967). Some experiments in comparative phonetics. In *International Congress of Phonetic Science*.
- Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., & Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5), 1526-1540.
- Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 14-23.
- Madsack, A. (2006). *Amplitude normalisation and intonation continuity modelling for unit selection* (Doctoral dissertation, Diplomarbeit, IMS, Universität Stuttgart, Stuttgart).

- Mahesha, P., & Vinod, D. S. (2012). Feature based classification of dysfluent and normal speech. In *Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology* (pp. 594-597). ACM.
- Makkook, M. (2007). *A multimodal sensor fusion architecture for audio-visual speech recognition*. (Master Thesis).
- Maris, Y. (1966). *The Malay sound system*. University of Malaya.
- Marciniak, T., Krzykowska, A., & Weychan, R. (2012). Speaker recognition based on telephone quality short Polish sequences with removed silence. *Przegląd Elektrotechniczny*, 88(6), 42-46.
- McLaren, M., Baker, B., Vogt, R., & Sridharan, S. (2009). Improved SVM speaker verification through data-driven background dataset collection. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (pp. 4041-4044). IEEE.
- Medeiros, H., Batista, F., Moniz, H., Trancoso, I., & Nunes, L. (2013). Comparing different machine learning approaches for disfluency structure detection in a corpus of university lectures. *José Paulo Leal Ricardo Rocha*, 259.
- Medeiros, H., Moniz, H., Batista, F., Trancoso, I., & Nunes, L. (2013). Disfluency detection based on prosodic features for university lectures. In *INTERSPEECH* (pp. 2629-2633).
- Meduri, S. S., Ananth, R., Johansson, S., & Sällberg, B. (2011). A Survey and Evaluation of Voice Activity Detection Algorithms.
- Mercier, G., Bigorgne, D., Miclet, L., Le Guennec, L., & Querre, M. (1989). Recognition of speaker-dependent continuous speech with KEAL. *IEE Proceedings I (Communications, Speech and Vision)*, 136(2), 145-154.
- Meseguer, N. A. (2009). Speech analysis for automatic speech recognition. *Norwegian University of Science and Technology, Master's Thesis*, 109.
- Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
- Moniz, H., Trancoso, I., & Mata, A. I. (2009). Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. In *INTERSPEECH* (pp. 1719-1722).
- Mohamad, U. S., Shamsuddin, S. M., & Mahmud, R. (2009). Endpoint Detection Enhancement for Speaker Dependent Recognition. *Asia-Pacific Journal of Information Technology and Multimedia*, 7(1).

- Murakami, Y., & Mizuguchi, K. (2010). Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics*, 26(15), 1841-1848.
- Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2), 171-184.
- Nie, K., & Zeng, F. G. (2004). Using neural network and principal component analysis to study vowel recognition with temporal envelope cues. *In Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE* (Vol. 2, pp. 4592-4595). IEEE.
- Nilsson, M., & Ejnarsson, M. (2002). Speech recognition using hidden markov model. *Department of Telecommunications and Speech Processing, Blekinge Institute of Technology*.
- Nunes, R., & Neves, L. (2006). *Filled pause modeling*. Tech. report, L2F-Spoken Language Systems Laboratory, INESC ID Lisbon, Portugal.
- O'Connell, D. C., & Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay?. *Journal of Psycholinguistic Research*, 34(6), 555-576.
- O'Connell, D. C., & Kowal, S. (2004). The history of research on the filled pause as evidence of The Written Language Bias in Linguistics (Linell, 1982). *Journal of Psycholinguistic Research*, 33(6), 459-474.
- O'Shaughnessy, D. (1992). Recognition of hesitations in spontaneous speech. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 521-524. San Francisco, CA. IEEE.
- Ogata, J., Goto, M., & Itou, K. (2009). The use of acoustically detected filled and silent pauses in spontaneous speech recognition. *In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (pp. 4305-4308). IEEE.
- Paja, W., & Wrzesień, M. (2013). Melanoma important features selection using random forest approach.
- Panda, M., Abraham, A., & Patra, M. R. (2010). Discriminative multinomial Naive Bayes for network intrusion detection. *In Information Assurance and Security (IAS), 2010 Sixth International Conference on* (pp. 5-10). IEEE.

- Papoulis, A. (1984). Bayes' theorem in statistics and Bayes' theorem in statistics (reexamined). *Probability, random variables, and stochastic processes*. 2nd ed. New York, NY: McGraw-Hill, 38-114.
- Peddinti, V., & Hermansky, H. (2013). Filter-bank optimization for Frequency Domain Linear Prediction. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 7102-7106). IEEE.
- Pérez, A., Larrañaga, P., & Inza, I. (2009). Bayesian classifiers based on kernel density estimation: Flexible classifiers. *International Journal of Approximate Reasoning*, 50(2), 341-362.
- Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of the Acoustical Society of America*, 116(4), 2338-2344.
- Peters, J. (2003). LM Studies on filled pauses in spontaneous medical dictation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers- Volume 2* (pp. 82-84). Association for Computational Linguistics.
- Pillai, S. (2006). Self-monitoring and self-repair in spontaneous speech. *A Biannual Publication on the Study of Language and Literature*, 8(2) (2006), 114-126.
- Planet, S., & Iriondo, I. (2012). Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. In *Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on* (pp. 1-6). IEEE.
- Proakis John, G., & Manolakis Dimitris, G. (1996). Digital Signal Processing, principles, algorithms, and applications. *Pentice Hall*.
- Rabiner, L. R. (1977). On the use of autocorrelation analysis for pitch detection. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(1), 24-33.
- Rabiner, L. R., & Sambur, M. R. (1975). An algorithm for determining the endpoints of isolated utterances. *Bell System Technical Journal*, 54(2), 297-315.
- Reddy, A. A., Chennupati, N., & Yegnanarayana, B. (2013). Syllable nuclei detection using perceptually significant features. In *Proc. of interspeech*.

- Revathi, A., & Venkataramani, Y. (2012). Speaker Independent Connected Digit Recognition Using VQ and HMM in Additive Noise Environment. In *Advances in Computer Science and Information Technology. Computer Science and Engineering* (pp. 393-402). Springer Berlin Heidelberg.
- Ricke, A. D. (2006). *Automatic Frame Length, Frame Overlap and Hidden Markov Model Topology for Speech Recognition of Animal Vocalizations*(Doctoral dissertation, Faculty of the Graduation School, Marquette University).
- Rimer, M. E. (2007). *Improving neural network classification training* (Doctoral dissertation, Brigham Young University).
- Rosenberg, A., & Hirschberg, J. (2006). On the correlation between energy and pitch accent in read English speech. In *INTERSPEECH*.
- Rosdi, F., & Ainon, R. N. (2008). Isolated malay speech recognition using Hidden Markov Models. In *Computer and Communication Engineering, 2008. ICCCE 2008. International Conference on* (pp. 721-725). IEEE.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* 9, 65–78.
- Ruppert, D., S. J. Sheather, and M. P. Wand. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 90, 1257–69.
- Karim, S. N., & dan Pustaka, D. B. (1995). *Malay grammar for academics and professionals*. Dewan Bahasa dan Pustaka, Kementerian Pendidikan Malaysia.
- Sakhnov, K., Verteletskaya, E., & Simak, B. (2009). Dynamical energy-based speech/silence detector for speech enhancement applications. In *Proceedings of the World Congress on Engineering* (Vol. 1, p. 2).
- Sanchis, A., Juan, A., & Vidal, E. (2012). A word-based Naïve Bayes classifier for confidence estimation in speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(2), 565-574.
- Schuller, B., Müller, R., Lang, M. K., & Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *INTERSPEECH* (pp. 805-808).
- Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New-York: John Wiley & Sons.

- Seman, N., Bakar, Z. A., & Bakar, N. A. (2010). Measuring the performance of isolated spoken Malay speech recognition using Multi-layer Neural Networks. In *Science and Social Research (CSSR), 2010 International Conference on* (pp. 182-186). IEEE.
- Shaneh, M., & Taheri, A. (2009). Voice command recognition system based on MFCC and VQ algorithms. *World Academy of Science, Engineering and Technology*, 57, 534-538.
- Sheppard, S. E. (2013). *Application of a Naïve Bayes Classifier to Assign Polyadenylation Sites from 3'End Deep Sequencing Data*. (Doctoral dissertation, Department of Molecular Medicine University of Massachusetts).
- Shreve, D. H. (1995). Signal processing for effective vibration analysis. *IRD Mechanalysis Inc., Columbus, OH, USA*.
- Sheather, S. J. and M. C. Jones. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society* 53, 683–690.
- Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*(Doctoral dissertation, University of California at Berkeley).
- Shriberg, E., Bates, R. A., & Stolcke, A. (1997). A prosody only decision-tree model for disfluency detection. In *Eurospeech* (Vol. 97, p. 23832386).
- Silverman, B. W. (1986). Kernel density estimation technique for statistics and data analysis. *Monographs on statistics and applied probability*, 26.
- Singh, B., Rani, V., & Mahajan, N. (2012). Preprocessing In ASR for Computer Machine Interaction with Humans: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(3), 396-399.
- Snell, R. C., & Milinazzo, F. (1993). Formant location from LPC analysis data. *IEEE Transactions on Speech and Audio Processing*, 1(2), 129-134.
- Solà-Soler, J., Fiz, J., Morera, J., & Jané, R. (2011). Bayes classification of snoring subjects with and without sleep apnea hypopnea syndrome, using a Kernel method. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE* (pp. 6071-6074). IEEE.
- Soria, D., Garibaldi, J. M., Ambrogi, F., Biganzoli, E. M., & Ellis, I. O. (2011). A 'non-parametric' version of the Naive Bayes classifier. *Knowledge-Based Systems*, 24(6), 775-784.

- Stolcke, A., & Shriberg, E. (1996). Statistical language modeling for speech disfluencies. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on* (Vol. 1, pp. 405-408). IEEE.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B* 36, 111–47.
- Stouten, F. (2008). *Feature extraction and event detection for automatic speech recognition* (Doctoral dissertation, Ghent University).
- Stouten, F., & Martens, J. P. (2003). A feature-based filled pause detection system for Dutch. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on* (pp. 309-314). IEEE.
- Stouten, F., Duchateau, J., Martens, J. P., & Wambacq, P. (2006). Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation. *Speech Communication*, 48(11), 1590-1606.
- Sunny, S., Peter, D., & Jacob, K. (2013). Combined Feature Extraction Techniques and Naive Bayes Classifier For Speech Recognition. *CS & IT-CSCP*, 155163.
- Taher, R. S., Jamil, N., Nordin, S., & Bahari, U. M. (2014). A new false peak elimination method for poor DNA gel images analysis. In *Intelligent Systems Design and Applications (ISDA), 2014 14th International Conference on* (pp. 180-186). IEEE.
- Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis*, 495, 518.
- Tanaka, H., Tokunaga, K., Uchino, E., & Suetake, N. (2014). Classification of Intravascular Ultrasound Signal by Kernel Density Estimation and Bayes Theorem for Identification of Coronary Plaque Tissue.
- Temko, A., Macho, D., & Nadeu, C. (2008). Fuzzy integral based information fusion for classification of highly confusable non-speech sounds. *Pattern Recognition*, 41(5), 1814-1823.
- Ting, H. N., Jasmy, Y., Hussain, S. S., & Cheah, E. L. (2001). Malay syllable recognition based on multilayer perceptron and dynamic time warping. In *Signal Processing and its Applications, Sixth International, Symposium on. 2001* (Vol. 2, pp. 743-744). IEEE.

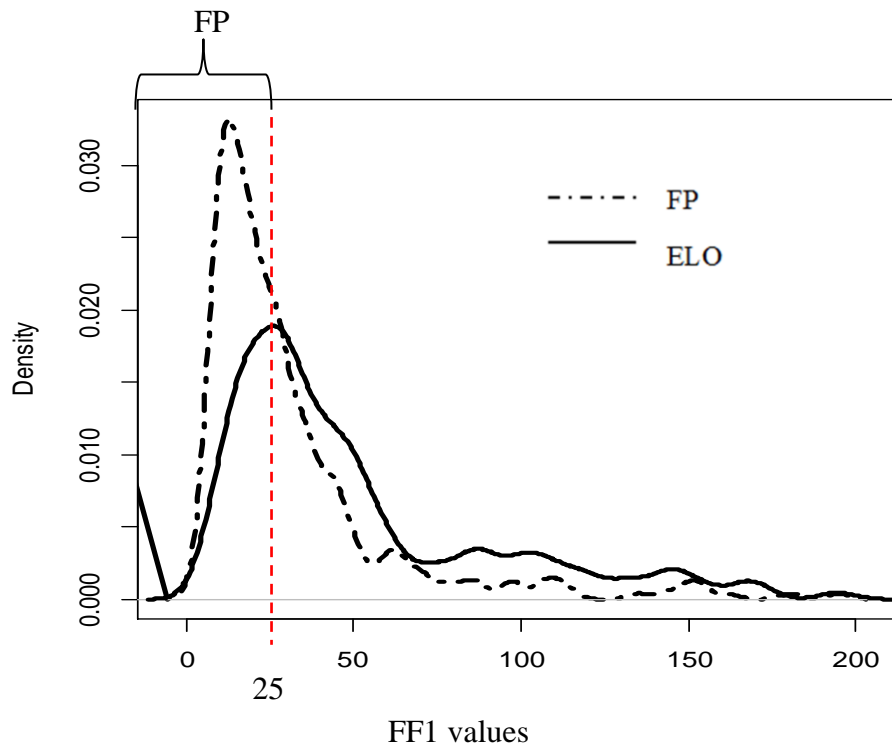
- Thomas, S., Ganapathy, S., & Hermansky, H. (2008). Recognition of reverberant speech using frequency domain linear prediction. *Signal Processing Letters, IEEE, 15*, 681-684.
- Trabelsi, I., & Ayed, D. B. (2013). A Multi Level Data Fusion Approach for Speaker Identification on Telephone Speech. *International Journal of Signal Processing, Image Processing and Pattern Recognition, 6*(2), 33-42.
- Veiga, A., Candeias, S., Lopes, C., & Perdigão, F. (2011). Characterization of hesitations using acoustic models. In *Proc. of the 17th International Congress of Phonetic Sciences, ICPhS XVII* (pp. 2054-2057).
- Verkhodanova, V., & Shapranov, V. (2014). Filled Pauses and Lengthenings Detection Based on the Acoustic Features for the Spontaneous Russian Speech. In *Speech and Computer* (pp. 227-234). Springer International Publishing.
- Vlasenko, B., Prylipko, D., & Wendemuth, A. (2012). Towards robust spontaneous speech recognition with emotional speech adapted acoustic models. In *Poster and Demo Track of the 35th German Conference on Artificial Intelligence, KI-2012, Saarbrücken, Germany* (pp. 103-107).
- Wang, B., & Zhang, S. (2005). A novel text classification algorithm based on Naïve bayes and KL-divergence. In *Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005. Sixth International Conference on* (pp. 913-915). IEEE.
- Womack, K., McCoy, W., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, P., & Haake, A. (2012). Disfluencies as extra-propositional indicators of cognitive processing. In *Proceedings of the workshop on extra-propositional aspects of meaning in computational linguistics* (pp. 1-9). Association for Computational Linguistics.
- Wu, C. H., & Yan, G. L. (2004). Acoustic feature analysis and discriminative modeling of filled pauses for spontaneous speech recognition. In *Real World Speech Processing* (pp. 17-30). Springer US.
- Xu, L., Krzyżak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *Systems, man and cybernetics, IEEE transactions on, 22*(3), 418-435.

- Yan, Q., Vaseghi, S., Zavarehei, E., Milner, B., Darch, J., White, P., & Andrianakis, I. (2007). Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing. *Computer Speech & Language*, 21(3), 543-561.
- Yang, Y., Wang, C., & Sun, Y. (2010). Speech recognition method based on weighed autoregressive HMM. In *Progress in Informatics and Computing (PIC), 2010 IEEE International Conference on* (Vol. 2, pp. 946-949). IEEE.
- Ye, J., Povinelli, R. J., & Johnson, M. T. (2002). Phoneme classification using Naive Bayes classifier in reconstructed phase space. In *Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop. Proceedings of 2002 IEEE 10th* (pp. 37-40). IEEE.
- Yokoyama, Z., Shinozaki, T., Iwano, K., & Furui, S. (2003). Unsupervised class-based language model adaptation for spontaneous speech recognition. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on* (Vol. 1, pp. I-236). IEEE.
- Yuan, L. (2008). An improved HMM speech recognition model. In *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on* (pp. 1311-1315). IEEE.
- Yusof, S. A. M., Paulraj, M., & Yaacob, S. (2009). Classification of Malaysian vowels using formant based feature. *Journal of ICT*, 7(2), 27-40.
- Zaidi, N. A., Cerquides, J., Carman, M. J., & Webb, G. I. (2013). Alleviating Naive Bayes attribute independence assumption by attribute weighting. *The Journal of Machine Learning Research*, 14(1), 1947-1988.
- Zambom, A. Z., & Dias, R. (2013). A Review of Kernel Density Estimation with Applications to Econometrics. *International Econometric Review (IER)*, 5(1), 20-42.
- Žgank, A., & Maučec, M. S. (2010). Modelling of Filled Pauses and Onomatopoeias for Spontaneous Speech Recognition. *SCIYO. COM*, 67.
- Zolnay, A., & Haeb-Umbach, U. D. I. R. (2006). *Acoustic feature combination for speech recognition* (Doctoral dissertation, RWTH Aachen University).
- Zuraidah, M. D., & Knowles, G. (2006). Prosody and turn-taking in Malay broadcast interviews. *Journal of Pragmatics*, 38(4), 490-5

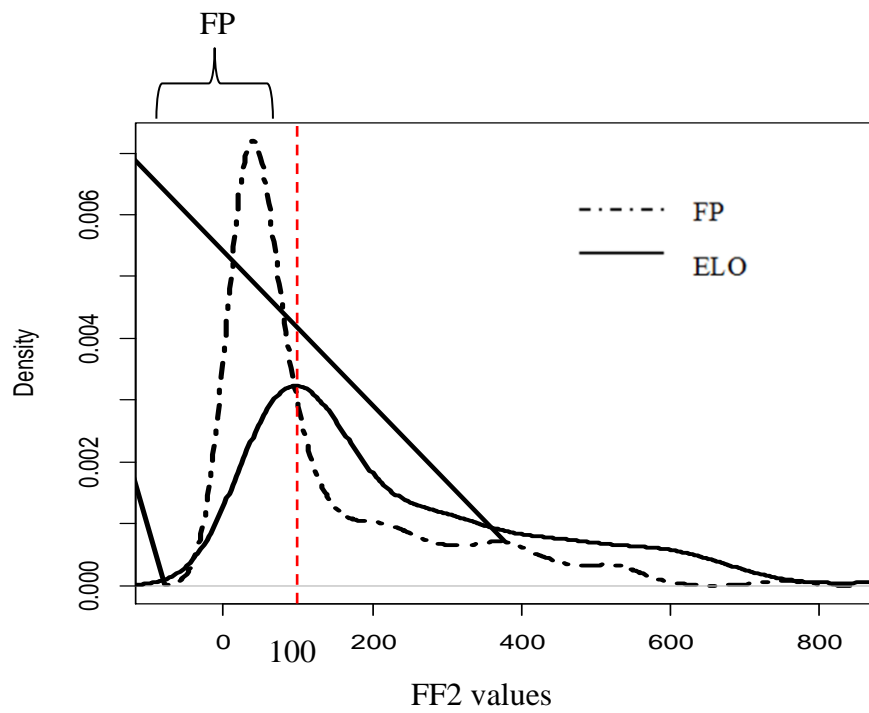
APPENDIX A

Kernel Density Estimation of Each Acoustical Feature

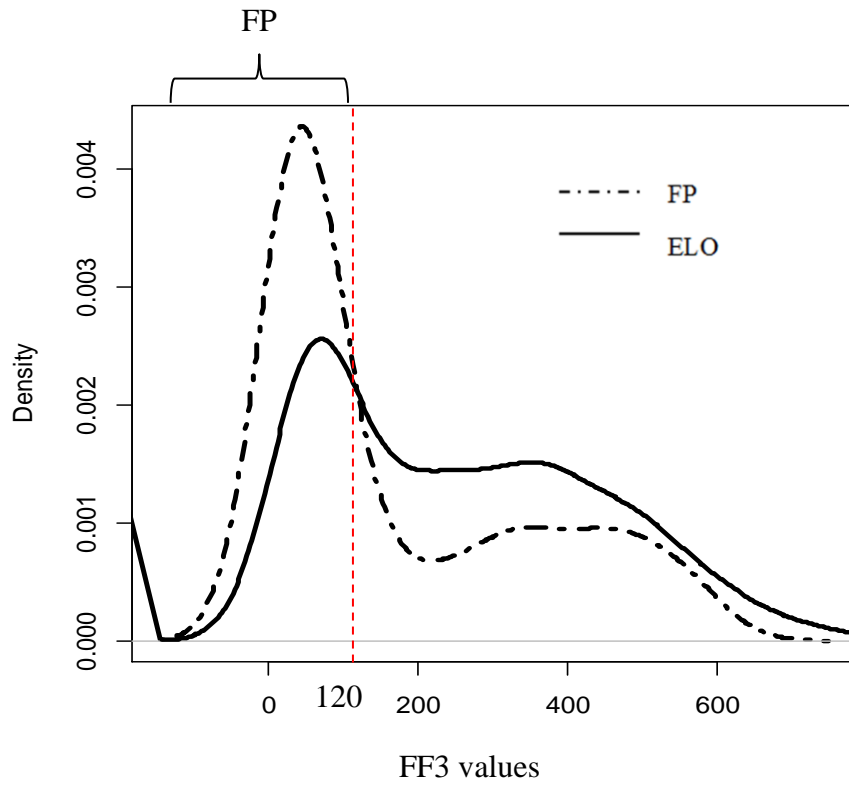
FF1 Kernel density



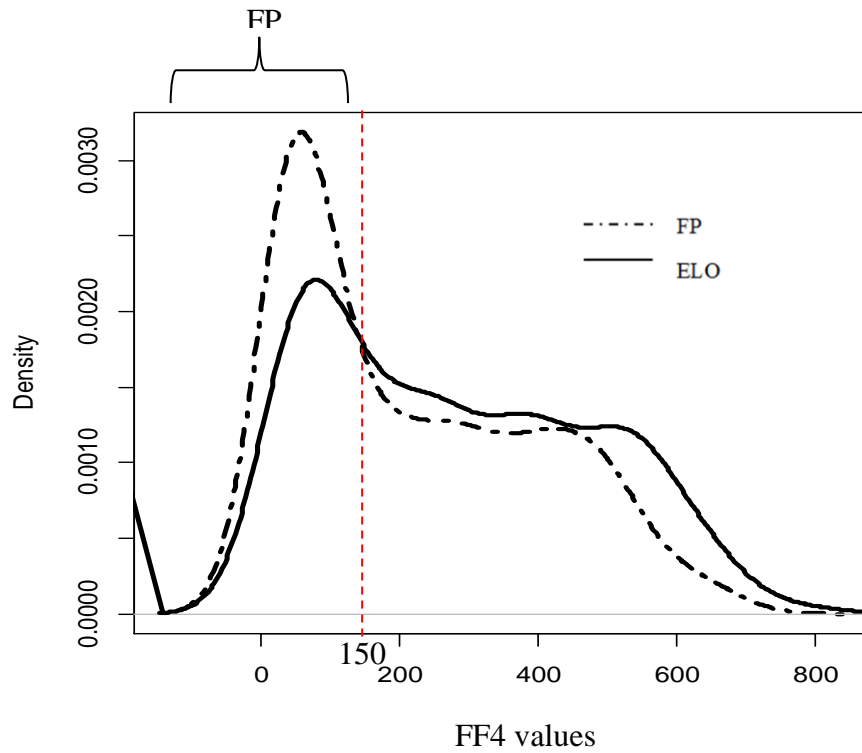
FF2 Kernel density



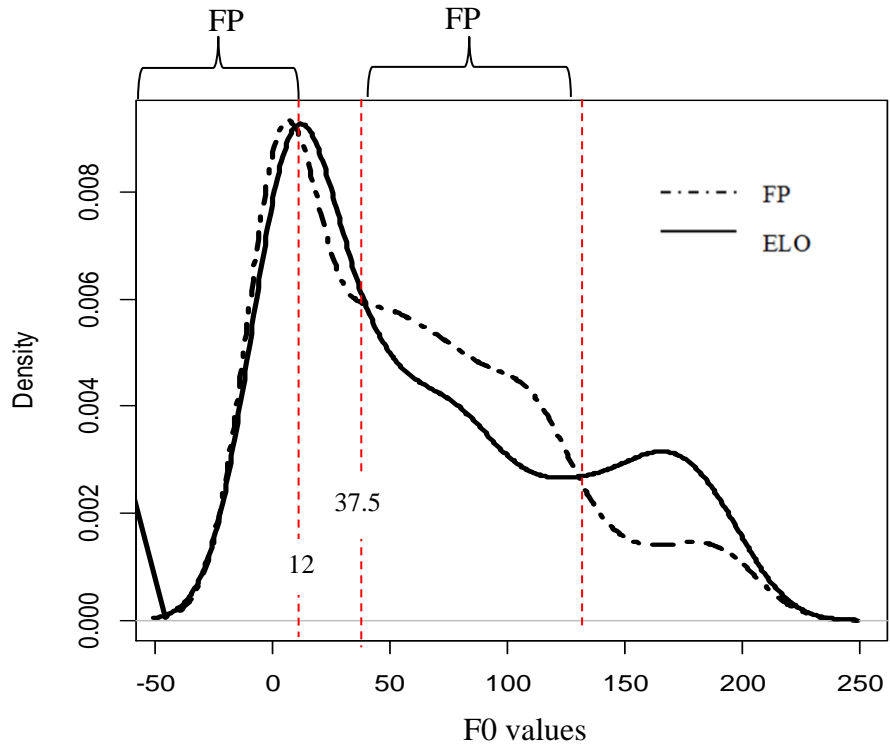
FF3 Kernel density



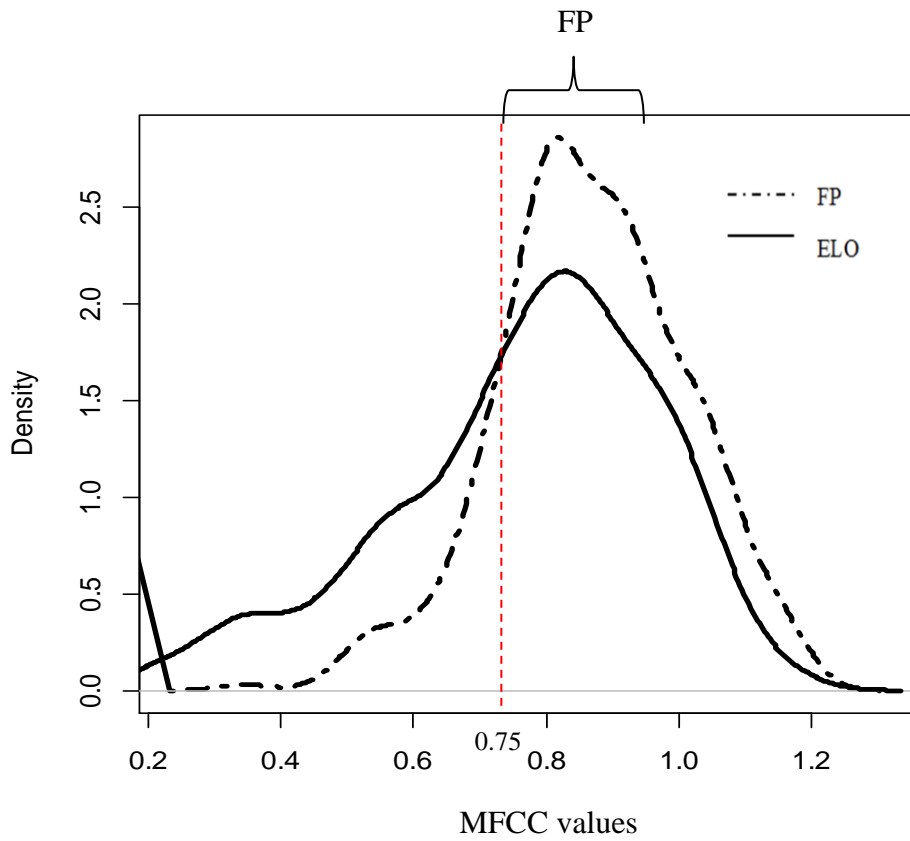
FF4 Kernel density



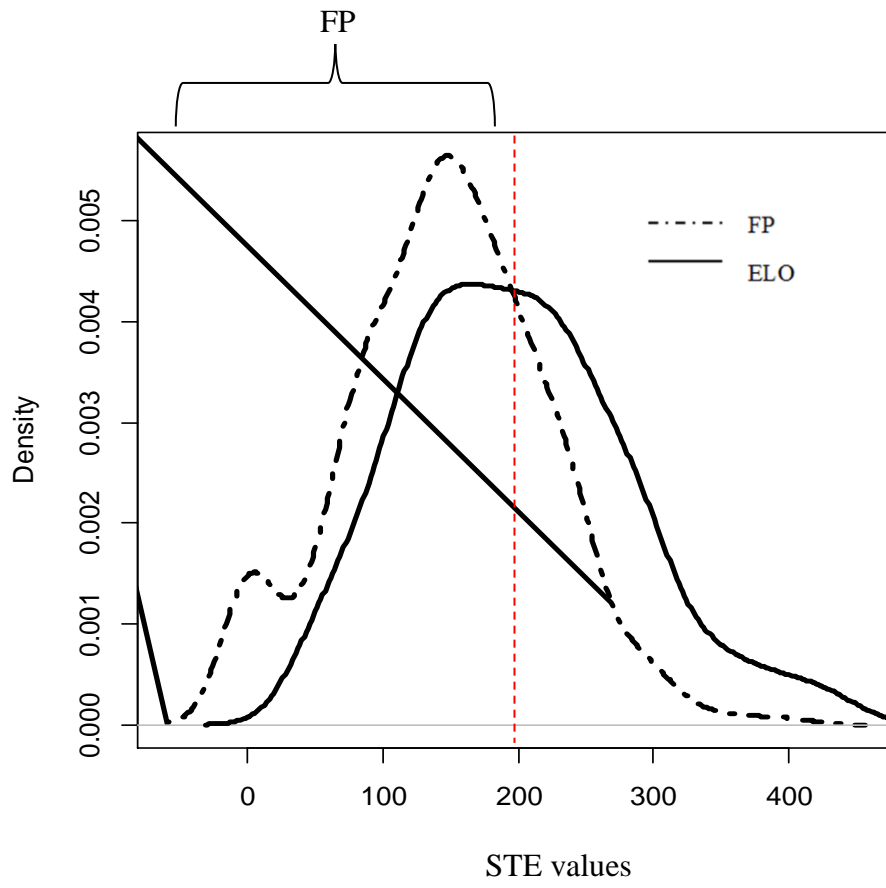
F0 Kernel density



MFCC Kernel density



STE Kernel density



APPENDIX B

Letter of Approval for Malay Language Filled Pause Sounds

1 Ogos 2013

Kepada Sesiapa Yang Berkenaan,

Tuan/Puan

PENGESAHAN KONSEP BUNYI FILLED PAUSE DAN ELONGATION BAHASA MELAYU

Dengan segala hormatnya perkara di atas dirujuk.

Saya seorang pensyarah bidang Linguistik Melayu dan telah menyemak data kajian berkaitan *Spontaneous Malay Language Disfluencies Classification*. Dengan ini saya mengesahkan bunyi-bunyi Filled Pause dan Elongation yang dikumpul daripada data kajian beliau, iaitu Malaysian Hansard Parliament Debate Database memang wujud dalam bahasa Melayu.

Sekian, terima kasih.

Yang benar,

Dr. Norizah binti Ardi

Ketua Pusat Pengajian Bahasa Melayu

Akademik Pengajian Bahasa

UiTM, Shah Alam.

AUTHOR'S PROFILE

Raseeda Binti Hamzah completed her PhD at the Faculty of Computer and Mathematical science, Universiti Teknologi MARA. She received her PhD in Information Technology and Quantitative Science, Universiti Teknologi MARA. She was a UiTM's Young Lecturer's Scheme from 2012-2015.

LIST OF PUBLICATIONS

1. Raseeda Hamzah, Nursuriati Jamil, Noraini Seman. **Acoustical Analysis of Filled Pause in Malay Spontaneous Speech**. In Tai-hoon Kim, Dae-sik Ko, Thanos Vasilakos, Adrian Stoica, Jemal Abawajy (eds.) Computer Applications for Communication, Networking and Digital Contents. Springer Series: Communications in Computer and Information Science, Springer-Verlag Berlin Heidelberg, vol. 350, pp. 251-259. 2012. ISBN: 978-3-642-35593-6. (**Best Paper Award**)
2. Hamzah, R., Jamil, N. & Seman, N. (2013). **Filled Pause Classification Using Energy-Boosted Mel-frequency Cepstrum Coefficients (MFCC)**. The 8th International Conference on Robotic, Vision, Signal Processing & Power Applications (RoVISP 2013), Penang, Malaysia, 10-12 November 2013.
3. Raseeda Hamzah, Nursuriati Jamil, Noraini Seman, **Impact of Acoustical Voice Activity Detection on Spontaneous Filled Pause Classification**, *5th IEEE conference on open system (ICOS 2014), Subang Jaya, 26-28 October, 2014*.
4. Seman, N., Jamil, N. & Hamzah, R. (2013). **Dynamic Connection Strategies (DyConS) for Spoken Malay Speech Recognition**. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2013), Athens, Greece, 12-15 December 2013.
5. Raseeda Hamzah, Nursuriati Jamil, Noraini Seman, **Nurturing Filled Pause Detection for Spontaneous Speech Retrieval, LNCS 2014**, 10th Asia Information Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014. Proceedings.