

**FUZZY PETRI NETS AS A CLASSIFICATION METHOD FOR
AUTOMATIC SPEECH INTELLIGIBILITY DETECTION OF
CHILDREN WITH SPEECH IMPAIRMENTS**

FADHILAH BINTI ROSDI

**FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2016

**FUZZY PETRI NETS AS A CLASSIFICATION
METHOD FOR AUTOMATIC SPEECH
INTELLIGIBILITY DETECTION OF CHILDREN WITH
SPEECH IMPAIRMENTS**

FADHILAH BINTI ROSDI

**THESIS SUBMITTED IN FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY**

**FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2016

UNIVERSITI MALAYA
ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Fadhilah Binti Rosdi

Registration/Matric No: WHA100021

Name of Degree: Doctor of Philosophy

Title of Thesis: Fuzzy Petri Nets as a classification method for automatic speech
intelligibility detection of children with speech impairments

Field of Study: Formal Methods

I do solemnly and sincerely declare that:

- i. I am the sole author/writer of this Work;
- ii. This Work is original;
- iii. Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
- iv. I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
- v. I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
- vi. I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature

Date: 22nd September 2016

Subscribed and solemnly declared before,

Witness's Signature

Date: 22nd September 2016

Name: Prof. Dr. Siti Salwah Binti Salim

Designation: Supervisor

Name: Dr. Mumtaz Begum Binti Peer Mustafa

Designation: Supervisor

ABSTRACT

The inability to speak fluently degrades the quality of life of many individuals. Early intervention from childhood can reduce disfluency of speech among adults. Traditionally, disfluency of speech among children is diagnosed based on speech intelligibility assessment by speech and language pathologists, which can be expensive and time consuming. Hence, numerous attempts were made to automate the speech intelligibility detection. While current detectors use statistical methods to discriminate unintelligible speech by calculating the posterior probability scores for each articulatory feature class, the major drawback is that the results are most likely to be based on training and input data, leading to inconsistencies in discriminating speech sounds. As such, the performance of detectors is below that of humans. To overcome this limitation, a new classification method based on Fuzzy Petri Net (FPN) is proposed to improve the classification accuracy. FPN was proposed as it has greater knowledge representation ability to reason using uncertain or ambiguous information. In this research, the speech features of Malay impaired children's speech are analysed for the identification of the significant speech features in the impaired speech which are related to the intelligibility deficits. This research also presents how the intelligibility classes can be detected by FPN. The results showed that FPN is more reliable in discriminating speech sounds than the baseline classifiers with improvements in the classification accuracy, precision and recall.

ABSTRAK

Ketidakupayaan untuk bercakap dengan fasih telah mengurangkan kualiti hidup bagi ramai individu. Intervensi awal ketika zaman kanak-kanak boleh mengurangkan masalah pertuturan di kalangan orang dewasa. Secara tradisinya, masalah pertuturan di kalangan kanak-kanak di diagnosis berdasarkan penilaian kejelasan pertuturan oleh pakar patologi yang melibatkan kos yang tinggi dengan tempoh yang agak lama. Banyak percubaan telah dibuat untuk mengautomasikan pengesanan kejelasan pertuturan. Walaupun pengesanan semasa yang menggunakan kaedah statistik boleh mendiskriminasi pertuturan yang tidak jelas dengan mengira skor kebarangkalian *posterior* bagi setiap kelas ciri pertuturan, ia bermasalah dalam menghasilkan keputusan yang paling mungkin berdasarkan latihan dan input data, yang membawa kepada percanggahan dalam ucapan membezakan bunyi. Oleh itu, prestasi pengesanan masih jauh dari apa yang pakar manusia mampu lakukan. Untuk mengatasi masalah ini, satu kaedah klasifikasi baru yang berasaskan Fuzzy Petri Nets (FPN) dicadangkan untuk meningkatkan ketepatan klasifikasi. FPN telah dicadangkan kerana ia mempunyai keupayaan perwakilan pengetahuan yang lebih baik dengan *reasoning* menggunakan maklumat tidak pasti atau samar-samar. Dalam kajian ini, ciri-ciri ucapan kanak-kanak yang bertutur dalam bahasa Melayu yang terjejas dianalisa untuk mengenalpasti ciri-ciri ucapan yang penting dalam ucapan individu bermasalah pertuturan yang berkaitan dengan defisit kejelasan. Kemudian, kajian ini membentangkan bagaimana kelas kejelasan boleh dikesan oleh FPN. Hasil kajian menunjukkan bahawa FPN adalah lebih dipercayai dalam membezakan bunyi pertuturan daripada *baseline classifiers* dengan peningkatan dalam klasifikasi ketepatan, ketepatan dan *recall*.

ACKNOWLEDGEMENTS

I would like to express deepest gratitude to my supervisors, Prof Dr. Siti Salwah and Dr. Mumtaz Begum for their full support, expert guidance, understanding and encouragement throughout my study. Without their patience and timely wisdom, I would never achieved this far.

I would also like to thanks Bassam Al Qattab and Adel Lahsasna for helping me with this research especially in the technical aspects. Thanks also goes to all my fellow friends in Human Computer Interaction (HCI) lab and Multimodal Interaction (MMI) lab, University of Malaya and all people involved directly and indirectly upon completing this research.

Last but not least, a very special thanks to my other half, Mohd Fazry Fabelilah who always there for me through thick and thin. My lovely princesses Shasmeen and Sarah who always bring happiness and kept me encouraged. My parents, En. Rosdi and Pn. Khadijah, parents in law, En. Fabelilah and Pn Siti Zubaidah, and siblings for their unconditional love and tireless support.

TABLE OF CONTENTS

Abstract	iv
Abstrak	v
Acknowledgements	vi
Table of Contents	vii
List of Figures	xiv
List of Tables	xvii
List of Symbols and Abbreviations	xx
List of Appendices	xxii
List of Publications	xxiii
CHAPTER 1 INTRODUCTION	1
1.1 Research Motivation.....	1
1.2 Research Background.....	3
1.2.1 Human Speech Production.....	3
1.2.1.1 Speech Sound and Features.....	4
1.2.2 Speech Impairments.....	6
1.2.2.1 Mispronunciations in Impaired Speech.....	9
1.2.3 Detection based ASR	10
1.3 Problem Statements.....	11
1.3.1 Issues with Speech Characteristics of Speech Impaired Speakers	11
1.3.2 Issues with the Intelligibility Detection System for Impaired Speech	13
1.3.3 Issues with the Classification Method in Detection System	14

1.4	Research Main Aim and Objectives	15
1.5	Research Questions (RQ)	16
1.6	Scope and Constraints	17
1.7	Research Methodology	18
CHAPTER 2 LITERATURE REVIEW		19
2.1	Speech intelligibility in Speech Impairments	19
2.1.1	Intelligibility Measurements	20
2.1.2	ASR and Speech Intelligibility	27
2.2	Automatic Speech Intelligibility Detection	28
2.2.1	Detection Strategies	29
2.3	Speech Corpus	32
2.3.1	Types of Speech	32
2.3.2	Size of Vocabulary	33
2.3.3	Speech Transcription	33
2.3.4	Available Speech Corpora for Impaired Speech	34
2.4	Feature Extraction	38
2.4.1	Prosodic Features	38
2.4.2	Pronunciation Features	40
2.4.3	Voice Quality based Features	44
2.4.4	Speech Features in Relation to Quality of Impaired Speech	46
2.4.5	Relationship between Speech Features and Speech Intelligibility	47
2.5	Classification Methods	48
2.5.1	Statistical Approach	49
2.5.2	Machine Learning Approach	51
2.5.3	Fuzzy Logic	53

2.5.4	Petri Nets	54
2.5.5	Fuzzy Petri Nets	54
2.5.6	Discussion	55
2.6	Available Classification Methods in Discriminating Impaired Speech Intelligibility	57
2.7	Summary	60

CHAPTER 3 RESEARCH METHODOLOGY 61

3.1	Findings of LR.....	61
3.1.1	Speech Corpus of Speech Impaired Children	61
3.1.2	Significant Speech Features	63
3.1.3	Speech Intelligibility Measurement	64
3.1.4	Speech Classification Method.....	64
3.2	Development of Children’s Impaired Speech Corpus.....	66
3.3	Selection of Significant Impaired Speech Features.....	70
3.4	Subjective and Automatic Measurement of Speech Intelligibility.....	71
3.5	Automatic Speech Intelligibility Detector.....	73
3.5.1	Speech feature extraction	74
3.5.2	Selection of Detection Strategies	75
3.5.3	The Baseline Classification Methods.....	75
3.6	The proposed FPN as a Classification Method in Automatic Speech Intelligibility Detection.....	76
3.6.1	Creating Fuzzy Inference System (FIS) using Subtractive Clustering	77
3.6.2	The Proposed FPN Classification	81
3.7	Evaluation of Classification Method	85

3.7.1	Accuracy, Precision and Recall.....	88
3.8	Summary	89

CHAPTER 4 A CORPUS OF MALAY SPEAKING CHILDREN WITH SPEECH IMPAIRMENTS: DEVELOPMENT AND ANALYSIS.....91

4.1	Speaker Characterization.....	91
4.2	Speech Materials and Stimuli.....	92
4.3	Recording Environment and Apparatus	96
4.4	Recording Procedure and Design	96
4.5	Reference Speech Corpus.....	98
4.6	Human Transcription and Labelling of the Impaired Speech Corpus.....	99
4.7	Speech Intelligibility Measurement.....	99
4.7.1	Subjective Evaluation by SLPs	100
4.7.2	Automatic Intelligibility Measurement using ASR	100
4.7.2.1	Data Preparation	100
4.7.2.2	Feature Extraction	104
4.7.2.3	Speech Training.....	104
4.7.2.4	ASR Evaluation.....	109
4.7.2.5	Result of ASR System.....	110
4.7.3	Discussion: Relationship between Intelligibility Scores and WRA	111
4.8	Speech Corpus.....	114
4.9	Analysis of Impaired and Control Speech in Relations to Intelligibility Deficits	115
4.9.1	Acoustic Analysis	115
4.9.2	Word Error Rate from ASR	116

4.9.3	Statistical Analysis	117
4.9.4	Results of Acoustic Analysis	117
4.9.5	Results of ASR performance.....	121
4.9.6	Summary of findings.....	121
4.10	Summary	123

CHAPTER 5 AUTOMATIC SPEECH INTELLIGIBILITY DETECTION FOR MALAY SPEECH IMPAIRED CHILDREN: A BASELINE RESULT 125

5.1	Automatic Speech Intelligibility Detection for Malay Impaired Speakers	125
5.1.1	Data Preparation.....	126
5.1.2	Speech Feature Extraction.....	126
5.1.3	Baseline Classification Methods	128
5.2	Evaluation of Baseline Classification Methods.....	128
5.3	Result.....	129
5.4	Discussion on Findings	133
5.4.1	Accuracy, Precision and Recall of Baseline Classification Methods	133
5.4.2	The relation of speech features with speech intelligibility classification.....	135
5.5	Summary	136

CHAPTER 6 THE PROPOSED FUZZY PETRI NETS (FPN) CLASSIFICATION METHOD FOR SPEECH INTELLIGIBILITY DETECTION

137

6.1	Proposed Approach	137
-----	-------------------------	-----

6.2	Fuzzy Inference System (FIS)	139
6.2.1	Fuzzy rule based reasoning and FPN classification	140
6.3	The FPN Implementation	142
6.3.1	The Petri Nets Modeling	143
6.3.2	PNML to GPenSIM files.....	144
6.3.3	GPenSIM Implementation	145
6.4	Evaluation of the Proposed FPN	146
6.5	Results	146
6.6	Discussion of Findings	149
6.6.1	Accuracy, Precision and Recall: FPN vs Baseline Methods.....	149
6.6.2	Misclassification and Error Rate.....	151
6.6.3	The Classification Accuracy of Speech Features: FPN vs Baseline Methods	151
6.6.4	Best Speech Features in Detecting Speech Intelligibility	153
6.7	Comparison with Existing Work.....	156
6.8	Summary	160
CHAPTER 7 CONCLUSION.....		161
7.1	Research objectives revisited	161
7.1.1	Research objective 1	161
7.1.2	Research objective 2	163
7.1.3	Research objective 3	164
7.1.4	Research objective 4	164
7.2	Research Contribution	165
7.3	Research Limitation	167
7.4	Future work	167

REFERENCES	170
Appendix A	181
Appendix B	184
Appendix C	185
Appendix D	186
Appendix E	188
Appendix F	194
Appendix G	262

University of Malaya

LIST OF FIGURES

Figure 1.1: Human vocal organs (Huang et al., 2001).....	3
Figure 1.2: The Human Speech Production Mechanism (Rabiner & Juang, 1993).....	4
Figure 1.3: Place of articulation (Huang et al., 2011).....	5
Figure 1.4: English vowel (Ladefoged & Maddieson, 1996).....	6
Figure 1.5: Three types of speech impairments	6
Figure 1.6: The example of mispronounced words.....	9
Figure 1.7: Detection based ASR (Bromberg et al, 2007).....	11
Figure 1.8: The illustration of the research scope	17
Figure 1.9: Research Phases.....	18
Figure 2.1: The Standard HMM based ASR system architecture (Jurafsky & Martin, 2009)	26
Figure 2.2: Block diagram of automatic speech detection (Canterla, 2012).....	29
Figure 2.3: Block diagram of a frame based detector (Canterla, 2012).....	30
Figure 2.4: Block diagram of a segment based detector (Canterla, 2012).....	31
Figure 2.5: The MFCC process (Jurafsky, 2009).....	41
Figure 2.6: A spectral slice from vowel [aa] before (a) and after (b) preemphasis (Jurafsky, 2009).....	41
Figure 2.7: The windowing process, where A is the frame size of 25mc and B is the frame shift of 10ms and a rectangular window (Jurafsky, 2009).....	42
Figure 3.1: The state of the art ASR system framework.....	73
Figure 3.2: Detection system framework	74
Figure 3.3: Integrating GPenSIM with the Fuzzy Logic Toolbox	84
Figure 4.1: Phoneme Distribution in the Speech Samples	95
Figure 4.2: Recording interface.....	96

Figure 4.3: Frequency of errors.....	99
Figure 4.4: The task grammar	101
Figure 4.5: The sample of pronunciation dictionary	102
Figure 4.6: A transcription at word level for the sentence “itu gajah”	103
Figure 4.7: A transcription at phoneme level for the sentence “itu gajah” (a) without a sp model (b) with a sp model.....	103
Figure 4.8: Speech training framework.....	105
Figure 4.9: List of the training files	106
Figure 4.10: Silence model	107
Figure 4.11: Example of conversion from (a) monophone model to (b) triphone model	108
Figure 4.12: Speech recognition framework.....	109
Figure 4.13: Sample of MLF transcription files.....	109
Figure 4.14: The correlations of the WRA with the intelligibility scores.....	112
Figure 4.15: Mean differences	121
Figure 4.16: Comparison of WER between SIG and CG	121
Figure 5.1: General framework of the speech intelligibility detection	126
Figure 5.2: Example of pitch contour of (a) intelligible speech and (b) not intelligible speech for sentence “itu gajah”	127
Figure 5.3: Comparison of the accuracy, precision and recall for baseline classification methods	134
Figure 5.4: The graph comparing the accuracy based on the prosody, pronunciation and voice quality	136
Figure 6.1: The framework of the proposed FPN	138
Figure 6.2: fismat structure	139
Figure 6.3: Fuzzy Inference System.....	140

Figure 6.4: Gaussian membership function plots.....	140
Figure 6.5: The FPN reasoning	142
Figure 6.6: The process flow of implementation	143
Figure 6.7: The PN Modelling using PN Editor	144
Figure 6.8: Integration of the MSF, PDF and TDF files (Davidrajuh, 2012)	145
Figure 6.9: Comparisons of the classification accuracy, precision and recall of FPN with the benchmark classifiers	150
Figure 6.10: The mean values of prosody, pronunciation and voice quality for FPN and baseline methods	153
Figure 6.11: Effect of (a) prosody, (b) pronunciation and (c) voice quality on classification accuracy	154
Figure 7.1 Mapping of the types of speech impairments and the aspect of speech features	162

LIST OF TABLES

Table 1.1.1: Summary of types in speech impairments	8
Table 2.1: The available measurement of speech intelligibility (Bauman-Waengler, 2012)	22
Table 2.2: Available researches on speech characteristics of impaired speech	28
Table 2.3: The Details of the Available Speech Impaired Corpora	35
Table 2.4: Relationship between the changes of speech features with speech quality ...	46
Table 2.5: Comparison of findings in speech features analysis of impaired speech.....	48
Table 2.6: Comparison of the approaches for the classification methods.....	55
Table 2.7: Summary of available speech intelligibility classification for impaired speech	59
Table 3.1: Malay consonants.....	67
Table 3.2: Malay vowels	68
Table 3.3: The speech transcribers' profile.....	68
Table 3.4: Phonological processes to describe phoneme changes (Hodson (1980); Ingram (1981); Shriber & Kwiakowski (1981)) Kahn (1982).....	69
Table 3.5: Differences between Mamdani and Sugeno type FIS (Reyes, 2012)	79
Table 3.6: A confusion matrix of classification results.....	85
Table 3.7: Summary of tasks, methods, approaches and more information on the implementation.....	89
Table 4.1: Sentences in the corpus with its IPA and SAMPA transcriptions	93
Table 4.2: Descriptions of the impaired speech corpus	97
Table 4.3: Number of male and female unimpaired speakers by age	98
Table 4.4: Intelligibility scores by SLPs	100
Table 4.5: Results of ASR baseline system	110

Table 4.6: Difference of subjective and automatic intelligibility scores	112
Table 4.7: Description of the speakers	114
Table 4.8: Selected sentences and words for vowels extraction	116
Table 4.9: The mean and s.d. of F1 and F2 for CG and SIG	118
Table 4.10: The mean and s.d. of F0 and intensity	118
Table 4.11: The mean and s.d. values of jitter and shimmer for the CG and SIG	119
Table 4.12: Differences between CG and SIG for each features	120
Table 4.13: Comparison of findings in acoustic analysis of impaired speech	122
Table 5.1: Confusion matrix of the baseline classification methods	129
Table 5.2: Classification accuracy of the baseline classification methods for each fold	130
Table 5.3: The overall classification accuracy, precision and recall of baseline classifiers	131
Table 5.4: The classification accuracy based on the individual speech features	131
Table 5.5: The degradation values of training and evaluation for each baseline methods	133
Table 5.6: Mean values for the accuracy, precision and recall	134
Table 5.7: The mean values of accuracy, precision and recall for each baseline methods	135
Table 6.1: Confusion matrix of the proposed FPN and the baseline classification methods	146
Table 6.2: Classification accuracy of the proposed FPN for all folds	147
Table 6.3: Classification accuracy, precision and recall of the proposed FPN	148
Table 6.4: The classification accuracy based on the individual speech features	148
Table 6.5: The mean values of accuracy, precision and recall for FPN and baseline methods	150
Table 6.6: Comparison of classification errors	151

Table 6.7: The classification accuracy of FPN and baseline classifiers for individual speech features	152
Table 6.8: The mean values for prosody, pronunciation and voice quality	153
Table 6.9: Correlation and coefficient of determination of prosody, pronunciation and voice quality	155
Table 6.10: Summary of related work on fuzzy petri nets for detection system	157
Table 7.1: Types of evaluation and their purposes	165

University of Malaya

LIST OF SYMBOLS AND ABBREVIATIONS

AI	:	Artificial Intelligence
CG	:	Control Group
CV	:	Consonant-Vowel
CVC	:	Consonant-Vowel-Consonant
DBASR	:	Detection based Automatic Speech Recognition
EDDM	:	Error Detection and Diagnosis Mechanism
FIS	:	Fuzzy Inference System
FPN	:	Fuzzy Petri Nets
HMM	:	Hidden Markov Model
HNR	:	Harmonic to Noise Ratio
HTK	:	Hidden Markov Model Toolkit
KNN	:	K-Nearest Neighbor
LDA	:	Linear Discriminant Analysis
LPC	:	Linear Prediction Coding
MFCC	;	Mel frequency cepstral coefficients
MSF	:	Main Simulation File
PCC	:	Percentage of Consonant Correct
PDF	:	Petri Net Definition File
PLP	:	Perceptual Linear Prediction
PN	:	Petri Nets
RF	:	Random Forest
SIG	:	Speech Impaired Group
SLP	:	Speech Language Pathology

SVM	:	Support Vector Machine
TDF	:	Transition Definition File
V	:	Vowel
VC	:	Vowel-Consonant
WER	:	Word Error Rate
WRA	:	Word Recognition Accuracy
ZCR	:	Zero Crossing Rate

University of Malaya

LIST OF APPENDICES

- Appendix A: Table 1- Speech Vocal Organ and Their Functions
- Table 2 - Description of the Place of Articulators
- Table 3 - Description of the manner of articulators
- Appendix B: Speech Intelligibility Assessment Form
- Appendix C: Speech Data for Classification
- Appendix D: Fuzzy Rules and Fuzzy Rules Viewer
- Appendix E: Iteration Count of FIS
- Appendix F: The pnml Document
- Appendix G: MSF, TDF and PDF files

LIST OF PUBLICATIONS

Journal

1. Mustafa M.B., Rosdi F., Salim S.S., Mughal M.U. (2015). Exploring the Influence of General and Specific Factors on the Recognition Accuracy of an ASR System for Dysarthric Speaker. *Expert Systems with Applications*, 42, pp. 3924-3932, 2015.
2. Rosdi F., Mustafa M.B., Salim S.S., B.A. Hamid. The Effect of Changes in Speech Features Towards the Recognition Accuracy of ASR System: A Study on the Malay Speech Impaired Children. *Malaysian Journal of Computer Science*. Accepted for publication

Conference Proceeding

1. Rosdi F., Mustafa M.B., Salim S.S. (2013). *The effect of speech rate on formant frequency movements*. The Annual International Conference on Science and Engineering in Biology, Medical and Public Health (BioMedicPub 2013).

CHAPTER 1 INTRODUCTION

This chapter provides research motivation and research background that lead to the derivation of the research problems, research aims and objectives as well as research questions. Research scope and constraints, research methodology and contributions are also being presented in this chapter.

1.1 Research Motivation

Speech is the sound produced by the interaction of human vocal organs. Speech is an important and vital mode of communication in our daily lives to express thoughts, ideas, emotions and convey messages to others. The inability to speak fluently can create problems for anyone. As a consequence, people might misunderstand the messages or even worse cannot be understood at all. On the other hand, the speaker is unable to convey or express their thoughts and ideas on what they meant, which can be very frustrating.

Disfluent speech, also known as mispronunciations, is a break, irregularities, or utterances that are often not consistent with any specific grammatical construction during smooth, meaningful flow of speech (Kates, 2008). Mispronunciations always occur in human speech. However, when a person's speech contains high frequency of mispronunciations, there is a high possibility that he or she suffers from speech impairment. Speech impairments affect the ability to produce speech. Basically, individual with speech impairment produces a speech pattern that differs from some standard pattern. It includes phrases, words, syllables or phonemes that are incorrectly uttered as well as incorrect or unintended intonation, stress and timing patterns (ASHA, 1993).

Intelligibility is referred as the accuracy of a listener in decoding the acoustic signal of a speaker (Yorkston et al., 1996). In other word, intelligibility is a measure of how much of an utterance can be understood. The concept of intelligibility is relevant to several fields such as clinical, phonetics and acoustical engineering. Assessing a person's intelligibility is highly relevant in clinical practice. According to Anumanchipalli et al. (2012), intelligibility assessment of pathological voices can be relevant both for diagnostic and therapy evaluation (Kim et al., 2015). Due to advancement of today's technology, there are numbers of detection applications developed to assess the speech intelligibility such as automatic intelligibility detection system. It has an important role in capturing atypical variation as well as opportune treatment of pathological voices (Huang et al., 2014).

Although there is a strong demand for accurate, reliable, and robust intelligibility assessment (Middag et al., 2009), the current state of the art system relies on the perceptual judgment of therapists, which is costly and time consuming.

Detection based Automatic Speech Recognition (DBASR) applications have shown to benefit people with speech impairments. Over the past decades, the developments of detection based ASR applications give invaluable contributions to the field of speech and language therapy in improving speech, language and communication skills among impaired speakers. However, the performance of speech detector is far inferior to human performance. The system performance is poorer when recognizing impaired speech due to its speech characteristics and intelligibility. There is a need for alternative way to reduce this performance gap. Therefore the motivation to conduct this research is to improve the ability of speech detector in discriminating the speech intelligibility for people who suffer from speech impairments. The proposed system will have the ability to improve the speech intelligibility by providing better detection of impaired speeches.

1.2 Research Background

This section provides the background of this research that consists of the following topics;

- Human speech production
- Speech impairments
- Detection based ASR

1.2.1 Human Speech Production

Speech is produced by the collaboration of human vocal organs. Figure 1.1 shows the human vocal organs such as lungs, vocal cords, tongue, nose and mouth that are responsible for producing speech.

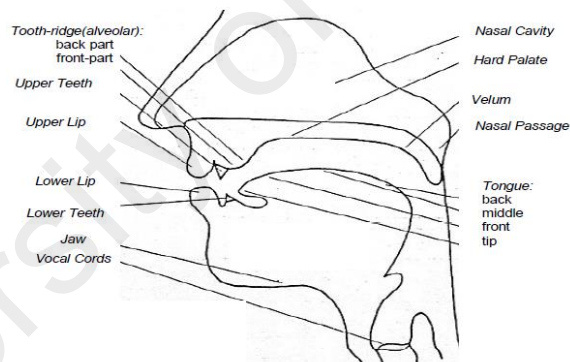


Figure 1.1: Human vocal organs (Huang et al., 2001)

Speech is produced as a sequence of sounds, while sound itself is produced by the rapid movement of air. Different sounds will be produced that depends on the state of the vocal cords, positions, shapes and sizes of the various articulators that changes over time (Rabiner & Juang, 1993). The vocal organs involve in the speech production mechanism are shown in *Table 1: Speech Vocal Organ and Their Functions (Appendix*

A) (Huang et al., 2001). Figure 1.2 shows the schematic diagram of the human speech production mechanism (Rabiner & Juang, 1993).

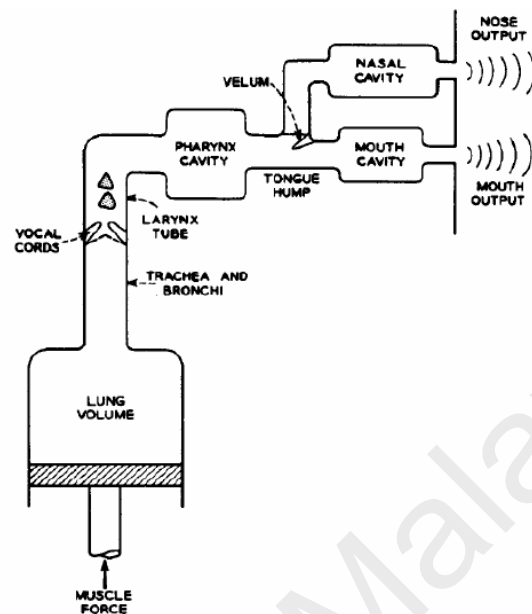


Figure 1.2: The Human Speech Production Mechanism (Rabiner & Juang, 1993)

1.2.1.1 Speech Sound and Features

Speech sound or phoneme are divided into 2 main classes; consonant and vowel. Consonants are produced through restriction or blocking of the airflow, which can be voiced or unvoiced. Vowels have less obstruction, usually voiced, and generally louder and longer than consonants.

- Consonants: Place of articulation and Manner of articulation

Consonants can be classified according to the place and manner of articulation. Place of articulation refers to the point of maximum restriction and can be distinguished by where the restriction is made (Jurafsky & Martin, 2007). These articulators are often used as a useful way of grouping phones together into equivalence classes (Jurafsky & Martin, 2007). Figure 1.3 shows the place of

articulation which consists of labial, dental, alveolar, palatal, velar and glottal. Each of the places of articulators are described in *Table 2: Description of the Place of Articulators (Appendix A)*.

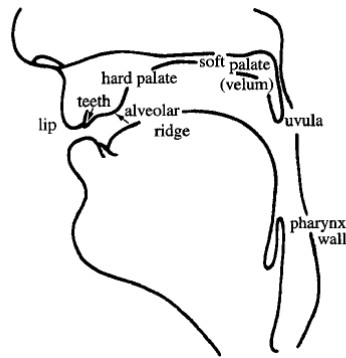


Figure 1.3: Place of articulation (Huang et al., 2011)

Manner of articulation is how the restriction in airflow is made and consists of stop or plosive, nasal, fricative, affricate, approximant, lateral and glide. Each of the manner of articulators are described in *Table 3: Description of the manner of articulators (Appendix A)*.

- Vowels

Vowels can be characterized by the articulator's position as they are made. The two most relevant parameters for vowels are called vowel height, which correlates with the location of the highest part of the tongue and the shape of the lips.

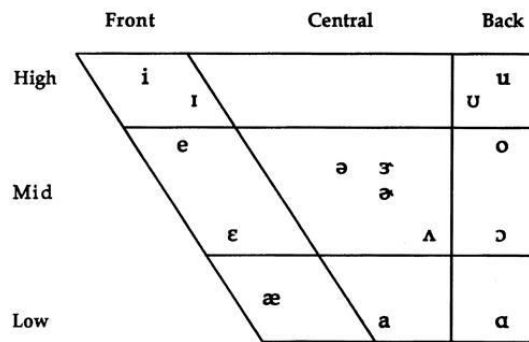


Figure 1.4: English vowel (Ladefoged & Maddieson, 1996)

1.2.2 Speech Impairments

According to the American Speech-Language-Hearing Association (ASHA) guidelines, speech impairment is used to indicate oral and verbal communication which is so deviant from the norm that it is noticeable or interferes with communication (ASHA, 1993). Speech impairments are categorized into three (3) basic types; 1) articulation disorders, 2) voice disorders and 3) fluency disorders as shown in Figure 1.5 (ASHA, 1993).

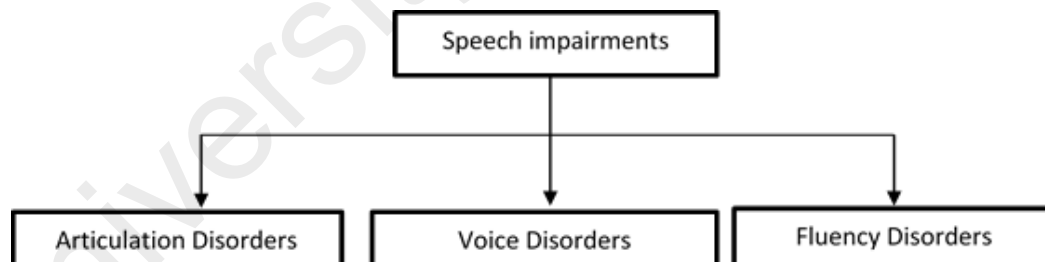


Figure 1.5: Three types of speech impairments

Articulation involves the gradual acquisition in moving the articulators in precise and rapid manner (Bauman-Waengler, 2012). In other words, articulation is a process of producing speech sounds that involve organs, manners and places of articulation. Thus, articulation disorder is the errors in the production of certain speech sounds

characterized by deletions or omissions, substitutions and distortions in the speech that degrade the speech intelligibility (ASHA, 1993).

Voice is produced by the vibration of the vocal cords. Air from the lungs sets these muscles into vibration, which is called phonation (Haynes et al., 2012). The voice is varied when it pass through the vocal cords, nose and mouth due to different size and shape spaces, which is called resonance (Haynes et al., 2012). Thus, voice disorders include aspect of phonation and resonance. A voice disorder refers to the abnormal production of speech properties like vocal quality, pitch, loudness, resonance, and/or duration (ASHA, 1993)

Fluency is the natural forward flow speech. A fluency disorder refers to the interruption in the flow of speaking, which may be accompanied by excessive tension, struggle behaviour, and secondary mannerisms (ASHA, 1993).

A person with speech impairments may have problem with articulation, voice or fluency or any combination of these. These impairments lead to the changes of the speech which affect the characteristics of the individual's speech. The speech characteristics of people with speech impairments vary depending on the type of impairment involved.

In many cases of speech impairments, *Dysarthria* and *Apraxia* are the examples of common articulation disorder resulted from motor impairment due to a disturbed neuromuscular control of speech mechanism (Kent et al., 1998) and disturbance in muscular movements (Darley et al., 1969) respectively. This physical limitation causes paralysis, weakness, or incoordination of the speech musculature (Darley et al., 1969; Nicolosi et al., 2004). Table 1.1 summarizes the types of speech impairments according

to the aspect of speech affected, speech characteristics and the example of disorders for each impairment types.

Table 1.1.1: Summary of types in speech impairments

Types of impairments	Aspect of speech affected	Speech Characteristics	Example
Articulation	Speech sound	Deletion <i>bo</i> for word <i>book</i>	Apraxia of Speech (AOS), Dysarthria
		Substitution <i>wabbit</i> for word <i>rabbit</i>	
		Distortion <i>Schit</i> for word <i>sit</i>	
Voice	Phonation	The voice may be harsh, hoarse, raspy, cut in and out, or show sudden changes in pitch with phonation disorders	Vocal nodules, Papilloma, Ulceration, Laryngeal web, Paralysis
	Resonance There are two (2) different types of resonance disorder: - Hyponasality - Hypernasality	<u>Hyponasality:</u> Insufficient voice energy from the nose, reducing the speech sound. <u>Hypernasality:</u> The movable, soft part of the palate (the velum) does not completely close off the nose, resulting in too much sound energy escapes through the nose.	<u>Hyponasality:</u> Blockage in the nose, Allergies <u>Hypernasality:</u> Cleft palate, Sub-mucous cleft, Short palate, Wide nasopharynx, Poor movement of the soft palate.
Fluency	Rhythm, Timing	An abnormal number of repetitions, hesitations, prolongations, or disturbances in the rhythm or flow	Stuttering
		Excessively fast and jerky speech	Cluttering

1.2.2.1 Mispronunciations in Impaired Speech

People with speech impairments suffer from the inability to pronounce phonemes properly that resulted in mispronunciations. There are four (4) types of common pronunciation errors that are identified in speech impairments;

- Substitution, *sub*, in which another phoneme, simpler to pronounce for the speaker is pronounced in the place of correct phoneme,
- Deletions, *del*, in which a phoneme is erased from the pronunciation of the speaker,
- Insertions, *ins*, in which an extra phoneme appears intercalated in the pronunciation,
- Distortions, *dis*, in which the phoneme is incorrectly pronounced, but it resembles another phoneme of the speaker's language.

Figure 1.6 below shows the word “*intention*” recognized as “*execution*”, where the aligned strings are a series of symbols that expresses an operation list for converting the top string into the bottom string; *del* for deletion, *sub* for substitution and *ins* for insertion.

i	n	T	e	*	n	t	i	o	n
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
	e	x	e	c	u	t	i	o	n
del	sub	sub		ins	sub				

Figure 1.6: The example of mispronounced words

1.2.3 Detection based ASR

A detector is basically a binary classifier that distinguishes between patterns with common quality (the class) and the rest (the anti-class) (Canterla, 2012). Sub-word detection in speech refers to the process of discovering segments of speech signal that belong to a given sub-word class. Automatic Speech Recognition (ASR) is the automated process in converting speech into text. Speech detection and recognition share similar problems; the former focuses on separating one speech class from the rest while the latter tries to separate every class from the others.

Detection of phonetic events such as phones and articulatory features has many applications such as computer aided pronunciation training (CAPT) and detection-based automatic speech recognition (DBASR). However ASR systems performance in recognizing speech is not equal to the performance of human. As such, there is a need to look for alternative structures that can reduce this performance gap.

It is important to emphasize that event detection is different from and harder than event recognition. Event detection in continuous time series involves both localization and recognition. Given a time series, a detector must localize the starts and the ends of target events and then recognize their classes. Event recognition systems, such as those from Yamato et al. (1992), Brand et al. (1997), Gorelick et al. (2007), Sminchisescu et al. (2005), and Laptev et al. (2008), only need to classify pre-segmented subsequences that correspond to coherent events. Because events are fundamental components of time series, event detection is an important problem. It is a cornerstone in many applications, from video surveillance (Piciarelli et al., 2008) and earthquake detection (Roberts et al.,

1989) to motion analysis (Aggarwal & Cai, 1999) and psychopathology assessment (Cohn et al., 2009).

Figure 1.7 depicts the framework of DBASR as proposed in Automatic Speech Attribute Transcription Project (ASAT) (Bromberg et al., 2007). The structure of a DBASR system basically consists of a bank of detectors and a linguistic merger. The bank of detectors analyzes the speech signal that belongs to a class. These detectors used classification methods to discriminate the speech features. Information from the detectors is processed by a linguistic merger and an output sequence of linguistic units is then hypothesized. Therefore, in DBASR, accurate detectors and classification methods are decisive for the performance of the system.

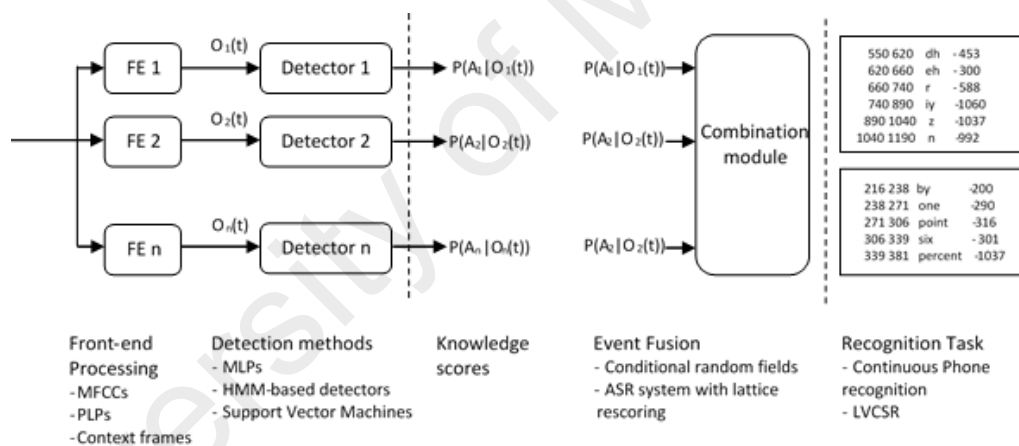


Figure 1.7: Detection based ASR (Bromberg et al, 2007)

1.3 Problem Statements

This section highlights the issues concerning problems that arise in the current intelligibility detection system for speech impaired speakers.

1.3.1 Issues with Speech Characteristics of Speech Impaired Speakers

Characteristics of impaired speech often related with disturbance and higher variability in speech. In most cases, the same speaker might produce different words

such as “fish”, “fees”, “ish“, “dish“ for word “fish“ that varies each time the speaker speaks. This inaccurate sound production resulted in speech variability. Blaney and Wilson, (2000) reported that increase of variability is highly correlated with severe impairment that leads to reduction in intelligibility. According to National Centre of Voice and Speech (NCVS), disturbances or changes in the output of the voice are commonly described by perturbation and fluctuation. A perturbation is generally identified as a small, temporary change in the vocal system while a fluctuation is a more significant change and tends to indicate that the voice is somehow unstable (NCVS). Several studies (Kent et al., 2000; Liss et al., 2002; Rosen et al., 2003; Ogawa et al., 2010) revealed that disturbances in speech impairments lead to intelligibility deficits in speech.

The impaired speech of children is different from adult, where the acoustic and linguistic characteristics of children’s speech are grossly different from adult speech (Shahin et al., 2015; Sztahó et al., 2014; Saz et al., 2009). For example, children’s speech is characterized by higher pitch and formants frequencies compared to adults’ speech. On top of that, children’s speech characteristics vary with age due to the anatomical and physiological changes during a child’s growth (Giuliani & Gerosa, 2006).

Children make mistakes when learning to utter new words. A speech sound disorder occurs when these mistakes continue past a certain age. Speech sound disorders include problems with articulation (making sounds) and phonological processes (sound patterns).

The intervention program for children is compelling as children learns better and faster than adults. However, not much progress was made in automatic speech detection

for children's speech as ASR based application is substantially more difficult for children's speech than for adults' speech (Russell et al., 2007). It is reported that the error rate of children are more serious than adults (Saz et al., 2009; Balter et al., 2005; Coleman & Meyers, 1991).

Some of the differences between children's and adults' speech due to physiological differences, namely children's vocal tracts are smaller than adults (Russell et al., 2007). These differences are caused by anatomical and morphological differences in the vocal-tract geometry, less precise control of the articulators and less refined ability to control supra-segmental aspects such as prosody.

It was found that important differences in the spectral characteristics of children voices compared to adults include higher fundamental and formant frequencies, and greater spectral variability (Eguchi & Hirsh, 1969; Kent, 1976; Lee et al., 1999). Parametric models for transforming vowel formant frequency of children to the adult space were considered in several researches (Goldstein, 1980; Martland et al., 1996; Potamianos & Narayanan, 2003).

1.3.2 Issues with the Intelligibility Detection System for Impaired Speech

Although many applications have been developed for speech impaired users, the detection process becomes more challenging. This is due to the characteristics of the impaired speech that leads to alteration of speech production, which resulted in confusability of sound patterns. Higher variations in impaired speech pronunciation typically produced hundreds of different possible phoneme classification for each sound. Handling more variations leads to confusability in phoneme classification and poor performance by the conventional detection system.

One of the challenges in classification is that most patterns belong to the anti-class (Canterla, 2012), which is even more challenging for impaired speech. The characteristics of impaired speech that resulted in speech variability have been widely studied in the literature. According to Kim et al. (2015), these variabilities pose challenges for human expert's assessment. Meanwhile, variability in speaker factors, such as gender, age, dialectal, native/non-native difference, makes automated system development even more challenging.

In detection based ASR system, the selection of features is important for discriminating speech. On the same note, speech errors can be discriminated from regular speech by adopting the appropriate features. However, not many researches that have investigated suitable features for error detection in impaired speech. The speech characteristic of impaired speech is grossly different from regular speech, thus making the existing speech features to be less effective in recognizing impaired speech. As the usual features were not found to be representative of impaired speech, new features must then be identified.

1.3.3 Issues with the Classification Method in Detection System

Automatic detection is an alternative way to incorporate the speech knowledge sources such as phones and articulatory features for the detection task. Current systems use statistical methods as detector by calculating the posterior probability scores for each articulatory feature classes. Although using statistical methods is useful, it suffers the drawback of only producing result that is most likely based on training and input data. The system may give the best solution from a list of possible choices available from the input data during training. To have a very efficient system, the training data must be large enough, which sometimes could be difficult especially for the impaired speakers. Statistical approach tests program based on random inputs, are basically

simple and efficient in term of implementation, but were unable to prove the correctness of the program. It may generate invalid invariants with a small probability that leads to ambiguities and inconsistencies when generating results.

The discrimination of speech features is also performed using Machine Learning (ML) algorithms. Currently there are several types of algorithms being applied for speech error detection. The current ML algorithms in speech error detections were based on many of the existing work within the data mining domain. Each of these algorithms require sufficient amount of data for making prediction or classification. However, for speech impairment, researchers have to work with very limited data especially for children with speech impairments. As such, many of the existing ML algorithms applied in speech error detection were unable to produce maximum output due to limited data. Thus, new algorithm that can work well with limited data needs to be identified.

1.4 Research Main Aim and Objectives

This research aimed to improve the discrimination ability in automatic speech intelligibility detection for speech impaired speakers. To achieve the research's aim, specific objectives have been identified as follows;

1. To identify significant impaired speech features that are salient for the performance of automatic speech intelligibility detection
2. To identify a suitable classification method to enhance the performance of automatic speech intelligibility detection for speech impaired speakers
3. To develop an automatic speech intelligibility detector based on the identified classification method in objective 2

4. To evaluate and compare the performance of the proposed classification method with existing baseline methods

1.5 Research Questions (RQ)

The specific research questions have been identified according to five research objectives as aforementioned.

Objective 1: To identify significant impaired speech features that is salient for the performance of automatic speech intelligibility detection

RQ1: What are the relevant speech features that could potentially affect the intelligibility of impaired speech?

RQ2: What speech features should be measured for the Malay pronunciation of speech impaired speakers in automatic speech intelligibility detection?

Objective 2: To identify a suitable classification method to enhance the performance of automatic speech intelligibility detection for speech impaired speakers

RQ3: What is needed to optimize the discrimination ability in the automatic speech intelligibility detection of impaired speech?

RQ4: What are the basis, structure and contents of the identified classification method?

Objective 3: To develop an automatic speech intelligibility detector based on the identified classification method in objective 2

RQ5: How will the proposed method developed?

Objective 4: To evaluate and compare the performance of the proposed classification method with the existing baseline methods

RQ6: What are the measurements used to evaluate the proposed methods?

RQ7: How the results of the proposed method being compared to the baseline method?

1.6 Scope and Constraints

This research focuses on improving the ability of detector realized with classification methods in order to classify the intelligibility speech in its class. As shown in Figure 1.8, this research involves several research areas from automatic speech recognition, speech pathology, and classification methods.

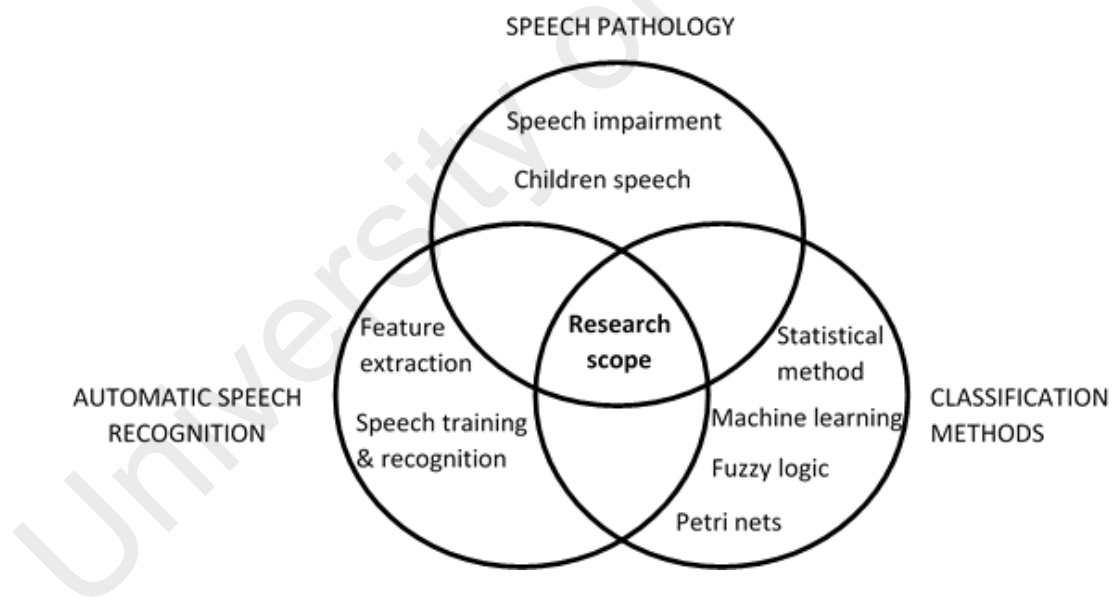


Figure 1.8: The illustration of the research scope

This research uses Malay speeches uttered by children with speech impairments. Children from different types of speech impairments are selected as sample populations

with age ranging from 8 to 12 years old for both genders. The selected children are from Klang Valley, Malaysia.

Methods and approaches in this work are specific to the structure of the Standard Malay language and the perception of speech impaired children. These methods and approaches could be applied to automatic speech intelligibility detection for other languages with appropriate and slight modifications with regards to the particular language. However, application to other languages is beyond the scope of this thesis.

1.7 Research Methodology

This research work is divided into three phases of study: Phase 1 is the Literature reviews work, Phase 2 is the Solution Construction, and Phase 3 is the Results Evaluation as shown in Figure 1.9. Each study is explained in detail in Chapter 3 on Research Methodology.

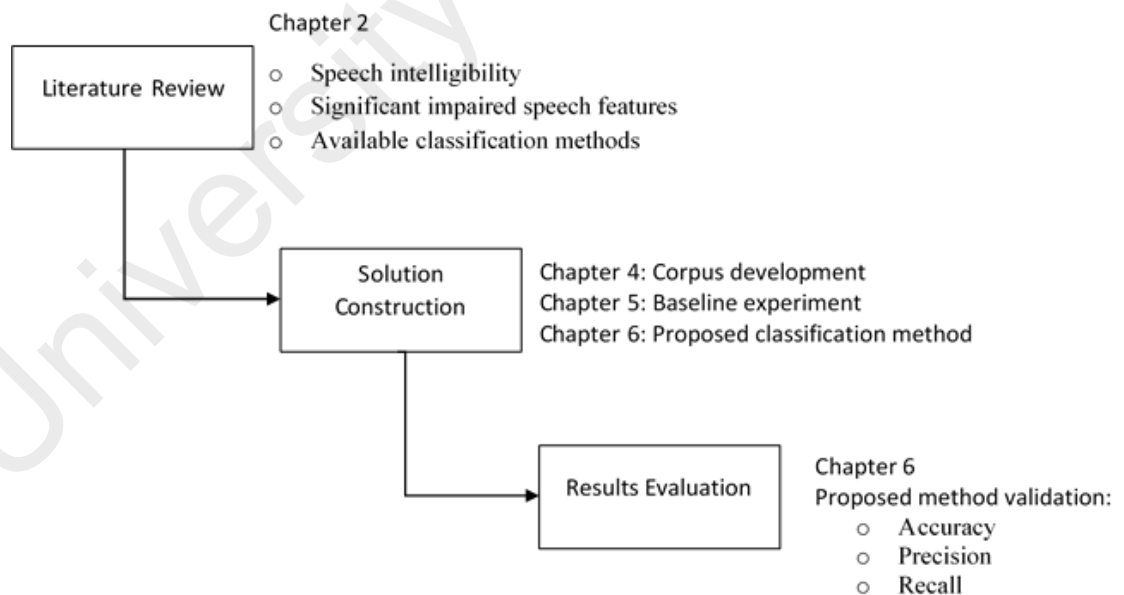


Figure 1.9: Research Phases

CHAPTER 2 LITERATURE REVIEW

This chapter addresses the main topics in this research; the intelligibility in speech impairments and speech detection systems. The first part reviews the characteristics of speech impairments and salient speech features in the detection system. The available databases for speech impaired speakers are also presented. The second part reviews the classification methods that includes the existing methods used in speech detection. Comparisons between the existing methods are also provided. Summary of the review is then discussed.

2.1 Speech intelligibility in Speech Impairments

Intelligibility refers to a judgement made by a clinician based on how much of an utterance can be understood (Bauman-Waengler, 2012). The speech characteristics of children with speech impairments are often found to be less intelligible than non-speech impaired children. The reduction in speech intelligibility is considered as one of the main characteristics of individuals with speech impairments. Intelligibility varies greatly depending on the extent of neurological disease or damage (Kent et al., 1989).

Speech impairments involve physical organ or articulators that are lacking articulatory precision. Speech signal varies due to differences in articulatory strategies across speakers (Wilson, 2004). For example, dysarthric speakers have high tendency to produce imprecise initial consonants for alveolar sounds such as d,n,s,t,z (Platt, 1980). According to Sawhney and Wheeler (1999), impaired speeches contain errors in place of articulation due to consonant confusions that related to alveolar such as labial (b/d, m/d) and velar (k/g, k/t) sounds. On the other hand, errors in the manner of articulation are due to consonant confusion pairs that related to fricatives or stops (f,p) and nasals (n,m) (Sawhney & Wheeler, 1999).

The jaw movement and uncontrolled breathing distress the consonant and vowel production due to the distortion in the formant pattern as well as the place and manner of articulation (Sy & Horowitz, 1993). Vowels are the easiest to produce physically because they do not require dynamic movement of the vocal system. However, phonetic transitions are the most difficult to produce because they require fine motor control to move the articulators precisely (Deller, 1991).

Several researches suggest that the correlation of speech variability and speech severity, where the greater speech variability, the greater is the severity of dysarthria (Ferrier et al., 1995; Doyle et al., 1997; Blaney & Wilson, 2000). The increasing severity of impaired speakers often correlates with decreasing of speech intelligibility (Ferrier et al., 1995; Doyle et al., 1997). Severity of speech impairments might differs among individual, but also differs for a single speaker due to several factors such as fatigue, stress as well as personal and environmental factors (Young, 2010). Intelligibility tests on impaired speakers produce most errors that are associated with phonemes that requires extreme articulatory positions such as stops (d,p,t) and fricatives (v,f,z,h) (Jayaram, 1995). In discriminating the speech sound, significant differences were found between vowels and consonants where the intelligibility for consonants was found to be substantially lower at 71% than vowels at 85% (Menendez, 1997).

2.1.1 Intelligibility Measurements

Speech intelligibility is a measurement that is commonly used to identify the severity level of impairment, which calculates the ratio of words understood by the listener to the total number of words articulated (Patel, 2002). There are three ways of measuring speech intelligibility as follows;

- Subjective measurement (Bauman-Waengler, 2012)

- Objective measurement (Bauman-Waengler, 2012)
- Automatic measurement (Schuster et. al, 2005)

For the subjective and objective measurement, the judgement is usually made by a clinician or speech language therapist (SLP) (Bauman-Waengler, 2012). On the other hand, the automatic measurement is made by computer based system such as the ASR (Schuster et al., 2005)

Subjective measurement

In subjective measurement, the measurements are based on a subjective and perceptual judgement that is generally related to the percentage of words that are understood by the listener (Bauman-Waengler, 2012). Factors that influence speech intelligibility include the number, type and consistency of speech sound errors (Bernthal et al., 2009). The number of errors is related to the overall intelligibility. However, just adding up the errors does not yield an adequate index of intelligibility. In Shriberg and Kwiatkowski (1982a, 1982b), a low correlation between the percentages of correct consonants with the intelligibility of speech were reported. Connolly (1986) listed factors that influence the intelligibility of utterances as follows;

- Loss of phonemic contrasts
- Loss of contrasts in specific linguistic contexts
- The number of meaning distinctions that are lost due to the lack of phonemic contrasts
- The difference between the target and its realization
- The consistency of the target realization relationship
- The frequency of abnormality in the client's speech
- The extent to which the listeners is familiar with the client's speech

- The communicative context in which the message occurs

Objective measurement

Although intelligibility is essentially a subjective evaluation, there were efforts to quantify it. There are several measures of intelligibility which includes indexing based on the frequency of occurrence of misarticulated sounds (Fudala, 2000), procedures that emphasize the phonetic contrast analysis (Monsen, 1981; Monsen et al., 1988; Ling, 1976; Kent et al., 1994), procedures that emphasize the phonological process analysis (Hodson & Paden, 1983; Leinonen-Davis, 1988; Webb & Duckett, 1990; Vihman & Greenlee, 1987), and procedures that emphasize the word level intelligibility (Yorkston & Beukelman, 1981; Wilcox et al., 1991; Weiss, 1982; Ingram & Ingram, 2001). Table 2.1 shows the existing measurement of speech intelligibility (Bauman-Waengler, 2012).

Table 2.1: The available measurement of speech intelligibility (Bauman-Waengler, 2012)

Measurement	Research	Procedure
Indexing based on the frequency of occurrence of misarticulated sounds	Fudala, 2000	Level 6: Sound errors are rarely noticed in continuous speech Level 5: Intelligible speech with noticeable errors Level 4: Intelligible speech with careful listening Level 3: Speech is regularly difficult Level 2: Speech is regularly unintelligible Level 1: Unintelligible speech
Phonetic contrast analysis	Monsen, 1981	CID Word Speech Intelligibility Evaluation (Word SPINE) for children and adults with severe and profound hearing impairments
	Monsen et al.,	CID Picture Speech Intelligibility Evaluation

	1988	(Picture SPINE) for children and adults with severe and profound hearing impairments
	Ling, 1976	Ling's Phonologic and Phonetic Level Speech Evaluation (PPLSE) for hearing impaired individual
	Kent et al., 1994	Children's Speech Intelligibility Test (CSIT) for children of any age
Phonological process	Hodson and Paden, 1983	Assessment of phonological Processes-Revised (APP-R) for children with object naming competence
	Leinonen-Davis, 1988	Function Loss (FLOSS) for children with phonological disorders
	Webb and Duckett, 1990	The RULES Phonological Evaluation for children with phonological disorders
	Vihman and Greenlee, 1987	Vihman-Greenlee Phonological Advance Measure for children especially those with phonological disorders
Word-level intelligibility	Yorkston and Beukelman, 1981	Assessment of intelligibility of Dysarthric Speech for adults and older children
	Wilcox et al., 1991	Preschool-Speech Intelligibility Measure (P-SIM) for preschool children and older children
	Weiss, 1982	Weiss Intelligibility Test (WIT) for children and adults
	Ingram and Ingram, 2001	Phonological Mean Length of Utterance (PMLU) and the Proportion of Whole-Word Proximity (PWP) for children primarily

Automatic measurement

Apart from the above procedures for objective measurements, there was also effort to use ASR system as a new approach for automatic measurements of speech intelligibility (Schuster et al., 2005). In Schuster et al. (2005), a modern word recognition system was developed and word recognition accuracy (WRA) was calculated. WRA is a standard measurement of recognizers evaluation that indicate how much a recognize word chain differ from the input speech. The calculation of WRA is in Equation 2.1:

$$\text{WRA (\%)} = ((\text{NC} - \text{NW})/\text{N}) * 100 \quad (2.1)$$

where NC is the number of correctly recognized words

NW is the number of wrongly inserted words

N is the total number of spoken words

In addition, the recognition result is also measured using the Word Error Rate (WER). WER is a standard calculation for recognizer evaluation. The calculation of WER is in Equation 2.2:

$$\text{WER} = \frac{I+S+D}{N} \quad (2.2)$$

where I is the total of insertion errors,

S is the total of substitution errors,

D is the total of deletion errors, and

N is the total number of all words in the transcription.

Standard Hidden Markov Model (HMM) based ASR System

A HMM is defined by 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). A standard HMM based ASR system typically consists of feature extraction, acoustic modelling and decoding phase as shown in Figure 2.1. Each of the components is described as follows;

- Feature Extraction

In feature extraction, the acoustic waveform is sampled into frames (usually 10, 15, or 20 milliseconds) which are then transformed into spectral features. It produces windows that are represented by a vector of 39 features that represents the spectral information, energy and spectral change.

- Acoustic Model

Acoustic Model provides statistical modelling to compute the likelihood of the observed spectral feature vectors given linguistic units or the acoustic observation sequence, O . The model units can be based on semantically meaningful units, such as words, or phonetically meaningful sub-word units such as phonemes.

- Language Model

Language Model provides linguistic and grammar constraints to the word sequence W which is often based on the statistical N-grams language models.

- Decoder

The decoding engine searches for the best phoneme sequence given the feature and the model. For this HMM based system, the Viterbi decoding is used as the decoding engine.

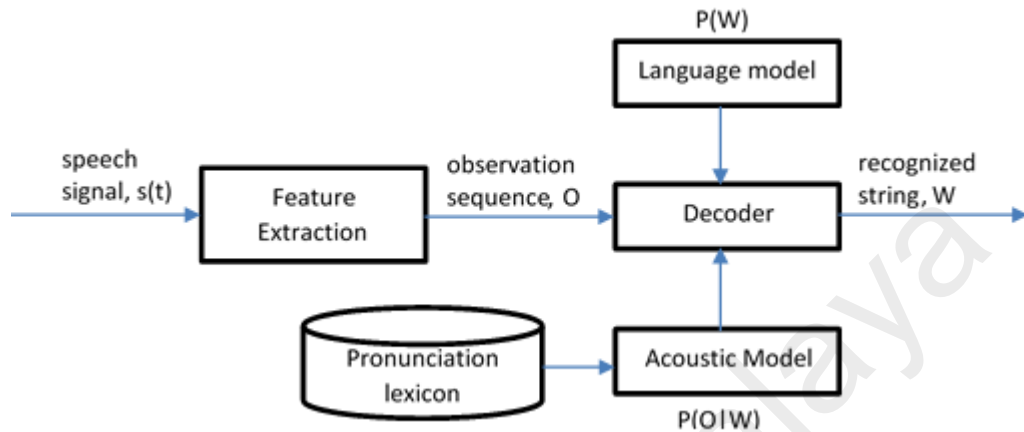


Figure 2.1: The Standard HMM based ASR system architecture (Jurafsky & Martin, 2009)

A speech signal $s(t)$ is input to ASR system that extract a set of speech features or observation sequence, O from the speech signal $s(t)$ in feature extraction. ASR system finds the most probable word, W given the observation sequence, O by taking the product of two probabilities for each word, and choosing the word for which this product is greatest (Jurafsky & Martin, 2009). The components of the HMMs speech recognizer which compute those two terms are; the $P(W)$, the prior probability computed by the language model and the $P(O|W)$, the observation likelihood is computed by the acoustic model as shown in Equation 2.3 (Jurafsky & Martin, 2009).

$$\hat{W} = \underset{W \in L}{\operatorname{argmax}} \underbrace{P(O|W)}_{\text{likelihood}} \underbrace{P(W)}_{\text{prior}} \quad (2.3)$$

To apply HMMs for speech recognition, three problems have to be overcome; the evaluation, the decoding and the training problem (Huang et al., 2001).

- The Evaluation Problem

The evaluation problem is to estimate the probability of observing the speech feature vector sequence given the HMM. The efficient solution is using the Forward and/or Backward algorithm.

- The Decoding Problem

The decoding problem is to find the best state sequence that is optimal in a certain sense given the speech feature sequence. The decoding problem is solved using the Viterbi algorithm.

- The Learning Problem

The learning problem is to estimate the HMM parameters from a given set of training samples according to some meaningful criterion. The best solution for this problem is using the Baum-Welch algorithm (Huang et al., 2001).

2.1.2 ASR and Speech Intelligibility

Ferrier et al. (1995) determined the correlation between speech intelligibility and characteristics, in relation to recognition accuracy, where low speech intelligibility leads to low recognition accuracy. There has also been growing interest to explore the speech characteristics of impaired speech in the motivation to develop ASR system that can improve the intelligibility of impaired speech. Rudzidc (2011), Kain et al. (2007) and Hosom et al. (2003) modified the speech features of dysarthria and reported that the modification of impaired speech shows increase in the ASR system recognition accuracy. Table 2.2 summarized the research works carried in the area of speech impairments and the ASR system performance.

Table 2.2: Available researches on speech characteristics of impaired speech

Reference	Language	Features	Findings
Ferrier et al., (1995)	English	Articulation, fluency and voice	Correlations between intelligibility measures and recognition success measures were strong where low intelligibility resulted in low recognition accuracy
Hosom et al., (2003)	English	Short-term spectral level: F0, Formant, Intensity	Short-term spectral level of dysarthric speech can be modified to improve intelligibility
Kain et al., 2007	English dysarthric speakers	F0, Formant, Intensity	Improving the intelligibility of dysarthric vowels of one speaker from 48% to 54%
Rudzic, 2011	English	Formant, F0	The correction of phoneme errors results in the increase of intelligibility in dysarthric speech

Though there are much effort in integrating additional speech knowledge into the ASR system such as the speech features in order to increase its performance, ASR systems still have limitation in classification task. Strik (2005), state that, the standard data driven approach to ASR system may not use all available knowledge about speech or language. One way to integrate the beneficial knowledge sources to ASR system in improving the classification performance is to extract knowledge based front-end features of Detection Based Automatic Speech Recognition (DBASR) as suggested by Li et al. (2005). There are many knowledge sources can be used such as phone or articulatory speech features that related to human speech production.

2.2 Automatic Speech Intelligibility Detection

Automatic speech intelligibility detection is one of the applications that make use of DBASR. In the context of intelligibility classification, the detector/classifier finds the

abnormal variation in the speech signal as unintelligible speech. Figure 2.2 depicts the intelligibility classification which consists of speech data, feature extraction and intelligibility classification.

Speech data or speech corpus consists of a collection of speech signals that are then extracted during the feature extraction, which produces the features that carry meaningful information for classification. In ASR system, feature extraction is common to all classes. On the other hand, there can be a specific feature extractor for each detector in classification task. This is an advantage because it is possible to process and extract relevant speech signals that are optimal for the specific class vs. anti-class problem in each detector (Canterla, 2012). The same speech features, however, are possible to be used in all detectors.

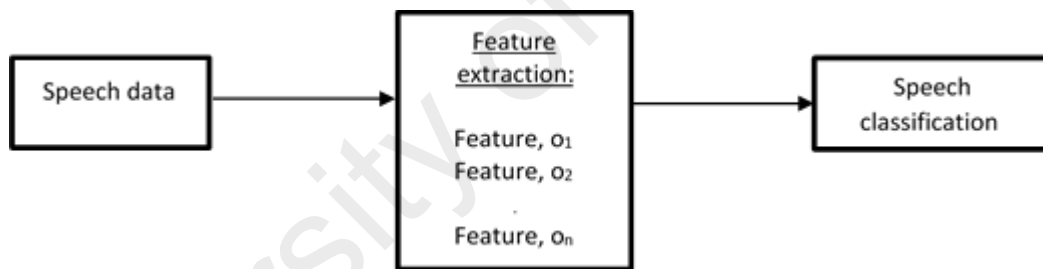


Figure 2.2: Block diagram of automatic speech detection (Canterla, 2012)

2.2.1 Detection Strategies

In this section, the design of sub-word detectors strategies are presented. There are two types of detection strategies which are frame-based and segment-based detectors where this categorization follows the work in (Li & Lee, 2005; Canterla, 2012).

Frame-based detection

Figure 2.3 shows a block diagram of the frame-based detector (Canterla, 2012) that classifies speech frames individually. The speech signal is input to the detector, which then produces a sequence of labeled frames. The first module is a feature extraction that extract discriminative information for the speech frames. Feature extraction is discussed further in Section 2.4.



Figure 2.3: Block diagram of a frame based detector (Canterla, 2012)

Next, the extracted features are input to a binary classifier that generates scores of the processed features. There are common strategies in generating scores of speech features. First is by generating a single class score for each frame. The second strategy is by generating scores for class and anti-class. Examples of common binary classifiers are Support Vector Machine (SVM), Gaussian mixture model (GMM), or Multi-Layer Perceptron (MLP). MLP is also known as artificial Neural Network (ANN). The classifier provides likelihood scores, which is further used in a decision rule for making a decision of the classification.

Segment-based detection

Figure 2.4 shows the block diagram of a segment-based detector that takes speech signal as an input and produces a sequence of labelled segments. The first module is the feature extractor, which is similar to the frame-based detector. However, for segment-based detection, all frames of the extracted speech feature are further processed to

identify the class segments. The segmentation is usually found with a decoding algorithm and statistical models for the class and the anti-class (Canterla, 2012).

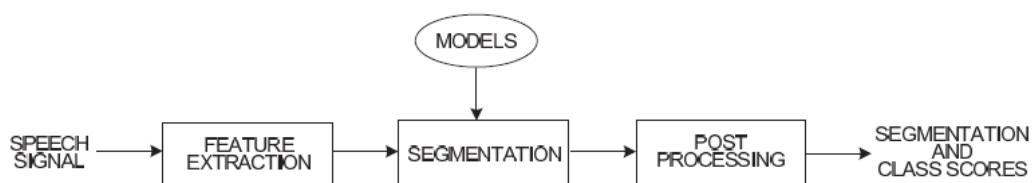


Figure 2.4: Block diagram of a segment based detector (Canterla, 2012)

Essentially, both frame and segment based detectors can be used. Li & Lee (2005) use frame based detectors with artificial neural networks (ANN) where the output scores can simulate the posteriori probabilities of an attribute given the speech signal (Li & Lee, 2005). In Hacıoglu et al. (2005), speech attribute detectors for manner and place of articulation were designed using ANNs with multiple outputs. These event detectors can also categorize each speech frame into one of the competing attributes. A good detector should only determine if the current speech frame exhibits the specified attribute or not. We need to group consecutive frames that have detection scores higher than a pre-selected threshold to form detected segments. It is clear that the frame detection scores are likely to fluctuate significantly, which results in extra detected segments. According to Li and Lee (2005), segment-based detectors were found to perform better than frame-based detectors. According to Nguyen (2012), the most popular approach is segment classification, which first selects candidate segments and then uses a classifier to predict if the segments belong to a target event class. Segment-based detectors can be combined with frame based detectors, or segment models such as hidden Markov models (HMMs), which have already been proved to be effective for ASR system.

2.3 Speech Corpus

Speech corpus is a collection of audio recordings of spoken language with transcriptions of the words spoken. In speech technology, a speech corpus or speech database is important for creating acoustic models to be used during the recognition task. It can also be used to analyze the speech characteristics and speech phonetics especially in the linguistic field.

One of the problems in DBASR is the lack of good speech corpus for impaired speech. The reason of database scarcity is due to the complexity of building a speech corpus for impaired speakers. In building a speech corpus, the process of speech acquisition is time consuming and involves many people, regardless the experimenter or the speakers (Saz, 2011). The acquisition process becomes more challenging for this research that involves children with speech impairments due to their severity level, physical and emotional state.

In Malay language, there is no existing impaired speech corpus for continuous speech. There was an effort for Malay speakers with speech impairments, specific to dysarthria (Al-Haddad, 2008) but limited to isolated words using digits from 0 to 9.

In developing a speech corpus, several aspects need to be considered such as the speech type, size of vocabulary and transcription type. These aspects will be further discussed in this section.

2.3.1 Types of Speech

In developing a speech corpus, two types of speech are usually considered (Anusuya & Katti, 2009);

- Speaking mode: isolated word, connected word, continuous speech

- Speaking style: dictation, spontaneous

ASR system can recognize isolated word, connected word or continuous speech. Isolated word contains single words or single utterance at a time. Connected word is similar to isolated words, but it allows separate utterances to be executed together at a time with a very minimum pause in between them. In continuous speech, users speak almost naturally and it is difficult to develop such system because of the special methods need to be utilized to determine the utterance boundaries (Anusuya & Katti, 2009). In speaking style, the speech can be either dictation or spontaneous. The dictation speech is a speech data recorded with speakers reading text provided to them while spontaneous speech is a natural speech without rehearsal.

2.3.2 Size of Vocabulary

The size of vocabulary affects the complexity, processing requirements and the accuracy of the ASR system (Hunt, 1997). Basically, size of vocabulary can be classified in three main classes; small, medium and large. A small size vocabulary usually contains less than 100 words (Campbell et al., 1999; Ashraf et al., 2010; Qiao et al., 2010). Example of ASR application using small size vocabulary are recognizing short instruction via telephone and recognizing command in portable device (Campbell et al., 1999). A medium size vocabulary corpus contains hundreds of words, from 100 and up to 1500 words. Meanwhile, large size vocabulary corpuses contain more than 1500 words and very large corpuses contain tens of thousands of words.

2.3.3 Speech Transcription

A speech transcription is prepared to represent distinct speech sound with a separate symbol, also known as phonetic transcription (Sharma, 2008). Phonetic transcription of a language is important because it contains information on how to pronounce a word

(Sharma, 2008). International Phonetic Alphabet (IPA) symbols and Speech Assessment Methods Phonetic Alphabet (SAMPA) are commonly used in phonetic transcription.

2.3.4 Available Speech Corpora for Impaired Speech

From 1993 to 2011, corpora of English impaired speeches have been developed such as the Whitaker database (Deller et al., 1993), the Nemours database (Menéndez-Pidal et al., 1996), Universal Access Database (Kim et al., 2008) and the more recent, TORGO database (Rudzicz et al., 2011). The development of speech corpora has become favourable from varying languages. There are several corpora of the impaired speech in languages other than English such as Alborada-13A for Spain (Saz et al., 2009), Mandarin (Jeng, 2006), Cantonese (Whitehill & Ciocca, 2000), Dutch (van der Molen et al., 2009) and Arabic (Attieh et al., 2010). CCM (Claude Chevré-Muller) and Aix-Neurology-Hospital corpus (ANH) are two corpora of French that contain large samples of speech data from French dysarthric speakers (Fougeron et al., 2010). In 2011, Korean Dysarthric Speech Database was developed to support the development of QoLT Software Technology. Table 2.3 summarizes the existing database.

Table 2.3: The Details of the Available Speech Impaired Corpuses

Reference	Language	No. of speakers	Speech type	Vocab size	Purpose
Whitaker database (Deller et al., 1993)	English	6 cerebral palsy speakers 1 healthy speaker	Isolated word	19,275 isolated word	Use in studies of recognition, perception, articulation, and other aspects of speech disorders.
Nemours database (Menéndez-Pidal et al., 1996)	English	<ul style="list-style-type: none"> • 11 male dysarthric speakers • 1 healthy speaker 	Continuous	<ul style="list-style-type: none"> • 814 short nonsense sentences • 74 sentences 	<ul style="list-style-type: none"> - To test the intelligibility of dysarthric speech - To investigate characteristics of dysarthric speech
Universal Access Database (Kim et al., 2008)	English	19 speakers with cerebral palsy (14 male and 5 female)	Isolated word	765 isolated word	For the impaired speech analysis, recognition and perception

NKI CCRT Speech Corpus (van der Molen et al., 2009)	Dutch	55 head and neck cancer patients	Sentence	935 sentences in each stage (total of 3 stages)	To study the speech intelligibility
TORGO database (Rudzicz et al., 2011)	English	7 speakers with cerebral palsy	Continuous	3,500 utterances in 23 hours of recording	To learn the articulatory features using computer speech models via statistical pattern recognition
Speech Database for QoLT Software Technology (Choi et al., 2011)	Korean	For dysarthric speech recognition • Dysarthric: 100 speakers • Healthy: 30 speakers	<ul style="list-style-type: none"> • Isolated word • Machine Control Commands • Korean Phonetic Alphabets, or Codes 	<ul style="list-style-type: none"> • Dysarthria: 35,900 utterances • Healthy: 17,850 utterances 	To develop the automatic assessment to access the degree of disability To investigate the phonetic features of dysarthric speech.
		For phonetic features of <u>dysarthric speech</u> • Dysarthric: 20 speakers	Isolated words	<ul style="list-style-type: none"> • Dysarthria: 4,200 utterances • Healthy: 	To study the phonetic characteristics of the different types of the disabled persons

		<ul style="list-style-type: none"> • Healthy: 10 speakers 		2,100 utterances	
CCM (Fougeron et al., 2010)	French	<ul style="list-style-type: none"> • 5000 speakers (adults and children) • 60 control speakers 	Word and sentence	1,000 hours of impaired speech	To develop a corpus of neurological speech disorders in monitor the evolution of dysarthria in a longitudinal French
ANH (Fougeron et al., 2010)	French	<ul style="list-style-type: none"> • 990 impaired speakers • 160 control speakers 	Word and sentence	Not available	To collect speech data from Parkinson's disease and Parkinsonian syndrome.

2.4 Feature Extraction

Feature extraction is the process of transforming the input speech waveform into a sequence of acoustic feature vector suitable for further speech processing. The objectives of this feature extraction are (Rosell, 2006):

- The features should extract the important aspects of the speech signal and should be perceptually meaningful.
- The features should be robust where the particular task should not be affected by the possible distortions, which can be caused by environmental and/or transmission medium.

There are several types of speech features. Basically, we can classify these features according to the aspect of prosodic, voice quality and pronunciation of pathological speech as proposed in Kim et al. (2015).

2.4.1 Prosodic Features

Prosody is the structure that organizes sound where tone, loudness, and the rhythm structures are the main components of prosody (Cutler et al., 1997). Suitable physical representations have to be formulated that include fundamental frequency or pitch, intensity, energy and the normalized duration of syllables. Intensity is the amount of energy that is transported past a given area of the medium per unit of time (Rosen & Howell, 2011). It correlates with the loudness of speech. On the other hand, fundamental frequency (F₀) is the lowest frequency that reflects the physiological limits of speech (Colton & Casper, 2006). Pitch is more closely related to the fundamental frequency where the higher the fundamental frequency is, the higher the pitch is perceived. However, discrimination between two pitches is depending on the lower pitch frequency. Perceived pitch will change when intensity is increased and frequency is kept constant (Huang et al., 2001).

The common energy related features are Signal energy and Zero Crossing Rate (ZCR). Signal energy is a time domain audio feature. The changes in energy is computed by dividing speech frames into sub frames of fixed duration. The normalized energy is divided with the total frame for each sub frame using Equation 2.4:

$$e_j^2 = \frac{E_{subFrame_j}}{E_{shortFrame_i}} \quad (2.4)$$

The total of all sub frames normalized energy is Energy entropy (EE) of that particular frame which is computed using Equation 2.5:

$$H(i) = -\sum_{j=1}^K e_j^2 \cdot \log(e_j^2) \quad (2.5)$$

Where $H(i)$ is EE of i th frame where e_j^2 is normalized energy of sub frames.

Zero crossing rate (ZCR) is also a time domain audio feature. ZCR is the rate of signal changes from positive to negative or back to its position, at the time signal have zero value. A ZCR is occurred in a signal when its waveform crosses the time axis or changes its algebraic sign which is computed using Equation 2.6:

$$Z_j = \frac{1}{2S} \sum_{i=1}^s \|sgn(x_i) - sgn(x_{i-1})\| \quad (2.6)$$

Where x_i is the discrete point of the i th frame and $sgn(.)$ is the sign function in Equation 2.7:

$$\text{Sgn}[x_i(n)] = \begin{cases} 1 & x_i(n) \geq 0 \\ -1 & x_i(n) < 0 \end{cases} \quad (2.7)$$

2.4.2 Pronunciation Features

Pronunciation features are spectral based features which usually represent the magnitude properties of speech spectrum (Jurafsky, 2009). It is commonly used in speech processing. In spectral related features, there are many possible feature representations such as LPC, PLP, Rasta and MFCC. By far the most common in the speech recognition is MFCC (Jurafsky, 2009). Formant frequencies are also common spectral features, which are the concentration of acoustic energy around a particular frequency in the speech wave (Lapteva, 2011). The formant with the lowest frequency is labelled as the first formant (F1), the higher is labelled as the second formant (F2), and the highest is the third formant (F3). These formants are closely related to the vowel production where F1 is related to the height of vowel, F2 is related to vowel frontness. F3 is considered to remain relatively constant for speakers (Nolan, 2002).

MFCC is capable to capture the important characteristic of audio signals. MFCC contains time and frequency information of the signal. MFCC has been widely used in the area of speech. The MFCC process is shown in Figure 2.5. The detail processes are further discussed in the following sub-section.

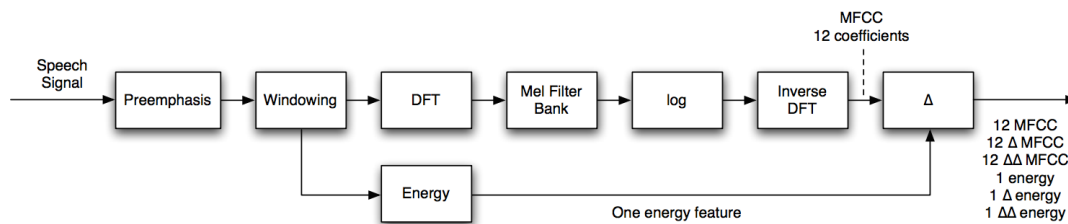


Figure 2.5: The MFCC process (Jurafsky, 2009)

Preemphasis

The preemphasis in MFCC feature extraction is to increase the amount of energy in the high frequencies. Increasing the high frequency energy makes information from these higher formants more available to the acoustic model and improves phone detection accuracy (Jurafsky, 2009). This preemphasis is done with a filter. Figure 2.6 shows an example of a spectral slice from the single vowel [aa] before and after preemphasis.

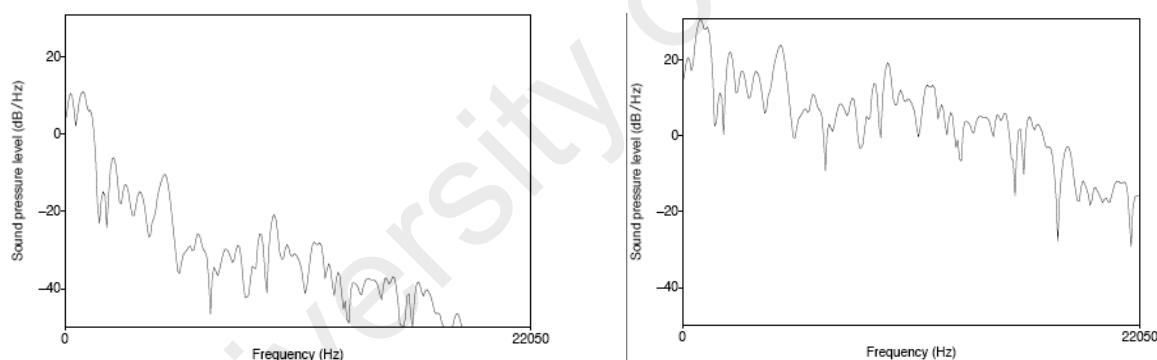


Figure 2.6: A spectral slice from vowel [aa] before (a) and after (b) preemphasis (Jurafsky, 2009)

Windowing

The goal of feature extraction is to provide spectral features that can be used to build phone or sub-phone classifiers. The spectral features are extracted from a small window of speech that characterizes a particular sub-phone where rough assumption made that the signal

is stationary. The extracted speech from each window is called a frame, the number of milliseconds in each frame is called a frame size and the number of milliseconds between the left edges of successive windows is the frame shift.

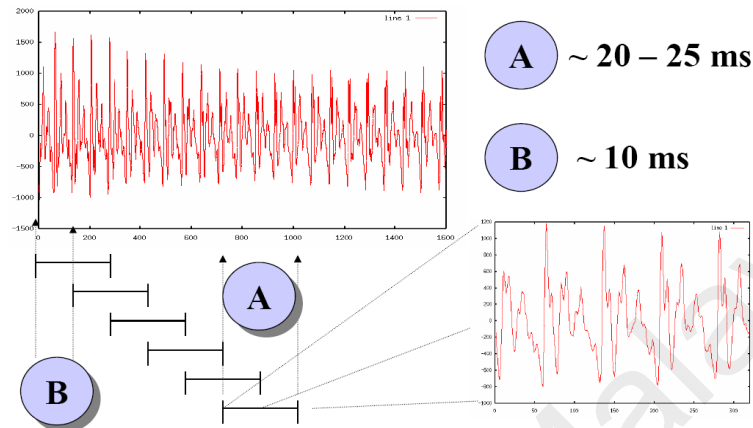


Figure 2.7: The windowing process, where A is the frame size of 25ms and B is the frame shift of 10ms and a rectangular window (Jurafsky, 2009)

The signal is extracted by multiplying the value of the signal at time n , $s[n]$ by the value of the window at time n , $w[n]$ using Equation 2.8:

$$y[n] = w[n]s[n] \quad (2.8)$$

Figure 2.7 shows the windowing process with the rectangular window shapes because the extracted windowed signal just like the original signal. The simplest window is the rectangular window which is derived from the Equation 2.9 below;

$$w[n] = \begin{cases} 1 & 0 \leq n \leq L - 1 \\ 0 & \text{otherwise} \end{cases} \quad (2.9)$$

However, the rectangular window can cause problems because it cuts off the signal at its boundaries which leads to discontinuities. It will be troublesome during the Fourier analysis.

Therefore, the Hamming window is used to avoid discontinuities by reducing the value of the signal towards zero at the boundaries. Therefore, the following Equation 2.10 is used;

$$w[n] = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{L-1}\right) & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases} \quad (2.10)$$

Discrete Fourier Transform (DFT)

A Fourier Transform is used to extract spectral information from the windowed signal which defined in Equation 2.11 as follows;

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\frac{\pi}{N}kn} \quad (2.11)$$

The windowed signals $x[n] \dots x[m]$ are the input to DFT and the output for each of N discrete frequency bands is a complex number $X[k]$ representing the magnitude against the frequency component in the original signal.

Mel filter Bank and Log

Filterbank analysis on the spectral representation of the speech signal is carried out by first creating a set of triangular filters on the mel-scale which is defined in Equation 2.12 as follows:

$$Mel(f) = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \quad (2.12)$$

where f is the frequency.

The Cepstrum (Inverse DFT)

To compute the cepstrum, it involves separating the source and the filter. The speech waveform is produced when a glottal source waveform of a particular fundamental frequency is passed through the vocal tract, which because of its shape has a particular filtering characteristic (Jurafsky, 2009). The most useful information for phone detection is the filter that the exact position of the vocal tract and cepstrum is one way to do this.

Deltas and energy

The extraction of the cepstrum with the inverse DFT produces 12 cepstral coefficients for each frame. Next, the energy from the frame is added as a 13th feature. Energy correlates with phone identity which is a useful cue for phone detection is computed in Equation 2.13 as follows:

$$Energy = \sum_{t=t_1}^{t_2} x^2[t] \quad (2.13)$$

The features related to changes in cepstral features over time are added by adding a delta and double delta features for each 13 features. Overall, 39 MFCC features derived which consists of 12 cepstral and 12 delta cepstral coefficients, 12 double delta cepstral coefficients, 1 energy coefficient, 1 delta energy coefficient and 1 double delta energy coefficient.

2.4.3 Voice Quality based Features

The voice quality based features is voicing related features that relate to the speech quality, with common features such as jitter and shimmer. Jitter and shimmer are acoustic characteristics of voice signals that are measured as the cycle-to-cycle variations of fundamental frequency and waveform amplitude, respectively (Farrús et al., 2007). Both

features correlate with the hoarseness in speech. There are several types of measurements for jitter and shimmer as follows (Vipperla et al., 2010);

- *Jitter (absolute)* is “the cycle-to-cycle variation of fundamental frequency”.
- *Jitter (relative)* is “the average absolute difference between consecutive periods, divided by the average period and expressed as a percentage”.
- *Jitter (rap)* is defined as “the Relative Average Perturbation which is the average absolute difference between a period and the average of it and its two neighbours, divided by the average period”.
- *Jitter (ppq5)* is “the five-point Period Perturbation Quotient, computed as the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period”.
- *Shimmer (dB)* is expressed as “the variability of the peak to-peak amplitude in decibels”.
- *Shimmer (relative)* is defined as “the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude which is expressed as percentage”.
- *Shimmer (apq3)* is “the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude”.
- *Shimmer (apq5)* is defined as “the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude”.

2.4.4 Speech Features in Relation to Quality of Impaired Speech

Several studies (Ikui et al., 2011; Niedzielska, 2001; Saz et al., 2009a; Wertzner et al., 2005; White, 2012) have investigated the characteristics and quality of speech features in speakers with speech impairments. It has been noted that speech quality has an effect on speech intelligibility (Amano-Kusumoto et al., 2014). Speech features of speech impairments change due to changes within the vocal organs (Niedzielska, 2001). Formant frequencies are characterized by the shape of the vocal tract. The vocal tract is always changing its shape during speech production, which leads to the change of the vowel quality (Ashley, 2013). Impaired speakers with problems to control the tongue movement will affect the formant values. Fundamental frequency (F0) and its harmonic components produced by vibration of vocal cords during speech production (Lemmetty, 1999). However, impaired speech affects vocal cords vibration (Darley et al., 1969). When the vocal cords are unable to move properly, it will produce voice problem, as well as breathing and swallowing problems (ASHA). Instability or lack of control in vocal cords vibration increases the jitter (Wilcox & Horii, 1980). On the other hand, shimmer is affected due to the glottis resistance reduction and mass lesions in the vocal folds (Wertzner et al., 2005). This produces creaky, hoarse and breathy sound, as well as limited pitch and loudness variations. Table 2.4 summaries the findings from literature related to the relationship between the changes of speech features with speech quality.

Table 2.4: Relationship between the changes of speech features with speech quality

Speech features	Changes in speech	Effect on speech quality
Formant	High / increase	Brighter sound (Parncutt & McPherson, 2002)
	Low / decrease	Darker sound (Parncutt & McPherson, 2002)
F0	High / increase	Louder sound (Lapteva, 2011)

	Low / decrease	Softer sound (Lapteva, 2011)
Jitter	High / increase	<ul style="list-style-type: none"> • Creaky, hoarseness in speech • Breathy sound • Rough sound (VAG)
	Low / decrease	<ul style="list-style-type: none"> • Smooth sound (VAG)
Shimmer	High / increase	<ul style="list-style-type: none"> • Decreasing voice loudness (Brockmann, 2011) • Hoarseness in speech • Softer voice
	Low / decrease	<ul style="list-style-type: none"> • Increase voice loudness (Brockmann, 2011) • Louder voice
Intensity	High	Intense, loud sound (Lapteva, 2011)
	Low / Decrease	Weak, soft sound (Lapteva, 2011)

2.4.5 Relationship between Speech Features and Speech Intelligibility

Significant changes in the speech features of impaired speech compared to normal speech have been found in previous studies (Hartl et al., 2003; Wertzner et al., 2005; Jeng et al., 2006; Saz et al., 2009). This section provides a review of those studies that examine which speech features contribute to speech intelligibility.

Among all features, the relationship between F0 and speech intelligibility has been investigated in several studies (Wertzner et al., 2005; Jeng et al., 2006; Saz et al., 2009; White, 2012). Jeng et al. (2006) reported that F0 is significantly correlated with sentence intelligibility. Wertzner et al. (2005) performed the intelligibility analyses of the vowel /a/ /e/ and /i/ and found that only F0 for /e/ was significantly correlated with intelligibility. Saz et al., (2009) and White (2012) found that there was no correlation between mean F0 and speech intelligibility. The question of whether F0 is an important cue for phoneme identification still remains debatable where some studies have reported that there is no significant difference in F0 decrement in impaired speech.

As for jitter and shimmer, Hartl et al. (2003) reported that these features are significantly correlated with intelligibility. However, these findings contradict those of other studies (Wertzner et al., 2005; White, 2012) which claimed that there are no differences in jitter and shimmer between impaired and non-impaired children. According to Wertzner et al., (2005) impaired speakers did not present any abnormality in the vocal folds, reduction of glottic resistance, vocal fold mass lesions and greater noise at production which may affect the values of jitter and shimmer. Table 2.5 summarizes the findings in the analysis of speech features in relation to intelligibility.

Table 2.5: Comparison of findings in speech features analysis of impaired speech

Author	Language	Features studied					
		Formant	F0	Intensity	Energy	Jitter	Shimmer
Saz et al., (2009)	Spanish	Significant	Not significant	Not significant	-	-	-
Jeng et al., (2006)	Mandarin	-	Significant	-	-	-	-
Wertzner et al., (2005)	Portuguese	-	Significant for vowel /e/	-	-	Not significant	Not significant
Hartl et al., (2003)	French	-	-	-	-	Significant	Significant
White, (2012)	English	-	Not significant	-	-	Not significant	Not significant

2.5 Classification Methods

Classification is a task of assigning an object that is characterized by a set of features or parameters to a class or category based on the characteristics similarities of the object. The classification task has been applied in a wide range of daily human activities such as, mechanical procedures in sorting letters based on the machine-read postcodes, assigning individuals to credit status based on their financial and personal information, and the

preliminary diagnosis of a disease to select immediate treatment for patients while awaiting definitive test results (Michie et al., 1994).

In the context of intelligibility discrimination, classification methods are used to classify speech intelligibility according to its classes as intelligible or not intelligible. In classifying speech intelligibility into its classes, the purpose is to understand the underlying processes that generate the classes of intelligibility and the occurrence of misclassification. Therefore suitable measures can be used to improve these processes. The most common approaches in classification methods are statistical approach and machine learning. However, fuzzy logic and formal learning approach Petri Nets and Fuzzy Petri Nets have been studied recently in the classification cases. In this sub-section, we discuss on the classification methods in terms of its usage, advantages and disadvantages.

2.5.1 Statistical Approach

In statistical approach, classification is also referred to as discrimination. According to An (2005), early work in classification focused on discriminant analysis which constructs a set of discriminant functions where linear functions of the predictor variables based on a set of training examples to discriminate among the groups defined by the class variable. Current studies discover more flexible model classes such as providing an estimate of the joint distribution of the features within each class using Bayesian classification, classifying an example based on distances in the feature space using the k-nearest neighbour method and constructing a classification tree that classifies examples based on tests on one or more predictor variables using classification tree analysis (An, 2005).

K-Nearest neighbour classifier (KNN)

According to An (2005), the KNN classifier classifies unknown examples to the most common class among its k nearest neighbours in the training data by assuming all the examples correspond to points in n -dimensional space. A neighbour is considered nearest if it has the smallest distance in the Euclidian in the n -dimensional feature space. The unknown example is classified into the class of its closest neighbour in the training set when $k=1$. KNN stores all the training examples and delays learning until a new example needs to be classified.

The advantage of KNN is intuitive, easy to implement and effective in practice (An, 2005). It can construct a different approximation to the target function for each new example to be classified, which is beneficial for complex target function (Mitchell, 1997). KNN produces good results if relevant features are used (White, 2000). However, the classifying cost for new examples can be high because almost all the computation is done at the classification time (An, 2005). The most serious shortcoming of KNN is that they are very sensitive to the presence of irrelevant parameters. Adding a single parameter that has a random value for all objects can cause these methods to fail (White, 2000).

Linear Discriminant Analysis (LDA)

LDA is commonly used as a dimensionality reduction technique in the pre-processing step for statistical and pattern-classification applications. The earliest work on LDA was described for a 2-class problem which was later generalized as multi-class Linear Discriminant Analysis or Multiple Discriminant Analysis (Raschka, 2014). The goal of an LDA is to build a feature space (a dataset n -dimensional samples) onto a smaller subspace k (where $k \leq n-1$) while maintaining the information of the class-discriminatory. In general, dimensionality reduction can reduce the computational costs for a given classification task and useful to

avoid overfitting by minimizing the error in parameter estimation, which is known as the curse of dimensionality (Raschka, 2014). The main advantage of LDA is the ability for dimension reduction of multiclass problem. Unlike other methods, LDA does not tackle a multiclass problem as a set of multiple binary class problems (Kim et al., 2007). LDA is easy to train with low variance. However, a limitation of LDA is that it needs at least one scatter matrix be nonsingular and breaks down when the data set is under sampled (Kim et al., 2007), and is bias if the model is incorrect.

2.5.2 Machine Learning Approach

Machine learning is generally used to encompass the automatic computing procedures based on logical or binary operations that learn a task from a series of examples (Michie et al., 1994). The goal is to generate the classification expressions that can be understood by humans. These techniques mimic the human reasoning into the learning process. The most commonly used machine learning is Decision Trees that learns the same tree structure as classification trees but uses a different criterion during the learning process (Michie et al., 1994). The other common methods in machine learning are Support Vector Machine and Artificial Neural Network.

Decision trees

Decision trees is a tree like graph classifiers. It is the most expressive and human readable representation of classification models (Mitchell, 1997). It uses a set of *if-then* rules where *if* condition1 and condition2 *then* outcome as in Equation 2.14:

$$\text{if } X > 1 \text{ and } Y = A, \text{ then } B \quad (2.14)$$

Rules can be generated in two ways (An, 2005); (1) translate a decision tree into a set of rules where one rule for each leaf node in the tree (2) learn rules directly from the training data. Decision trees perform faster than neural network methods in the training and application phase. The disadvantage is that they are not flexible enough at modelling parameter with complex distributions as compared to neural networks or KNN.

Random Forest (RF)

RF is a collaborative of decision trees. In standard trees, each node is split using the best split among all variables. In RF, each node is split using the best among a subset of predictors randomly chosen at that node. This counterintuitive strategy performs very well as compared to many other classifiers such as LDA, SVM, and ANN. In addition, RF is not sensitive to their values and efficient against overfitting (Breiman, 2001). However, RF predicts model using black box approach where it is difficult to interpret the prediction.

Support Vector Machine (SVM)

SVM is a discriminative classifier that finds a hyperplane to separate the d-dimensional data perfectly into its two classes (Vapnik, 1995). However, almost all data is not linearly separable, SVM's introduce the notion of a "kernel induced feature space" which casts the data into a higher dimensional space where the data is separable (Boswell, 2002). Typically, this would cause problems of computationally and overfitting. The advantage of SVM is that the higher-dimensional space doesn't need to be dealt with directly. Furthermore, the Vapnik-Chervonenkis (VC) dimension that measure a system's likelihood perform well on unseen data which can be explicitly calculated unlike other learning methods such as neural networks. Overall, SVM is intuitive, theoretically well- founded, and have shown to be practically successful (Boswell, 2002). However, SVM output is an uncalibrated class

membership probability, which is not a posterior probability of an input data. Another drawbacks of SVM is that the parameters of a solved model are difficult to interpret and not easy to incorporate prior knowledge into the model (Thuy et al., 2009).

Artificial Neural Networks (ANN)

ANN is one of the most widely known classifier. The main advantage of ANN is that it can handle problems with many parameters and are able to classify objects even when the parameter of features are very complex (White, 1996). However, the implementation of ANN is very slow in the training and the application. According to White (1996), another significant drawback of ANN is that it is very difficult to determine how the network makes its decision. It is difficult to determine the features that are important and useful for classification. Therefore, ANN has limitations in making decision of the best features that are important in developing a good classifier.

2.5.3 Fuzzy Logic

Fuzzy classification is the process of classifying objects or elements into a fuzzy set whose membership function is defined by the truth value of a fuzzy propositional function. It has been used in wide range of problem domains such as decision making, pattern recognition and classification. According to Nedeljkovic (2002), a fuzzy set is a set whose elements have degrees of membership where an element of a fuzzy set can be full member or a partial member. The membership value is assigned to an element that is no longer restricted to just two values such as yes/no or 0/1, but can be 0, 1 or any value in-between. Therefore, notions like rather tall or very small can be measured and processed by computer in order to apply the human way of thinking in the computer program (Hellman, 2001). Mathematical function that defines the degree of an element's membership in a fuzzy set is called membership

function. It uses the minmax rule for conjunctive (AND) and disjunctive (OR) reasoning where it takes the minimum and maximum of the membership functions (Perner & Petrou, 2003). However, minmax rule is not the way of human reasoning. It is possible there is enough training data in the learning process. Therefore, it chooses the best rule that fits the way of reasoning of the expert. Another disadvantage of the rules is that they give the same importance to all factors that are to be combined where this issue can be resolved if we do not insist on all membership functions taking values between 0 and 1 (Perner & Petrou, 2003).

2.5.4 Petri Nets

A Petri net is a graphical and mathematical modelling tool. The graphical representation consists of places (S), transitions (T), and arcs (F) that connecting the places and transitions. Transitions are active components that model activities which can occur when the transition fires that changes the state of the system. Transitions are only allowed to fire if all the preconditions for the activity are fulfilled where there are enough tokens available in the input places. When the transition fires, it removes tokens from its input places and adds some at all of its output places. The number of tokens removed or added depends on the cardinality of each arc. Petri nets have the ability to show a precise and graphical representation. It can be used as a visual-communication aid similar to flow charts, block diagrams, and networks. However, in Petri Nets, the token and transition are Boolean with restricted numerical values where, if the expression is true = 1 otherwise false = 0 (Hoheisal & Alt, 2007), and may be inadequate to address the problem of uncertainty and imprecision data due to the increasing of the system complexity in real world (Meher Taj & Kumaravel, 2015).

2.5.5 Fuzzy Petri Nets

Fuzzy Petri nets is widely used in many applications such as error detection and diagnosis mechanism (EDDM) in complex fault-tolerant PC-controlled system (Ting et al., 2008),

knowledge representation and reasoning (Li, 2000; Ribaric & Pavesic, 2009), and decision support system (Suraj, 2012). FPN is an integration of Fuzzy logic and Petri nets. Similarly, FPN has the ability to be combined with other approaches or tools such as Artificial Intelligence (AI) and mathematical models to become more efficient, and powerful (Aziz, et al., 2010).

2.5.6 Discussion

Table 2.6 summarizes the advantages and disadvantages of the existing classification methods.

Table 2.6: Comparison of the approaches for the classification methods

Methods	Advantages	Disadvantages
k-Nearest Neighbour	<ul style="list-style-type: none"> • Easy to implement • Easily understood by human • Proven high performance 	<ul style="list-style-type: none"> • Rather slow in training with many examples • Very sensitive to the presence of irrelevant parameters
Linear Discriminant Analysis	<ul style="list-style-type: none"> • Simple to use • Dimensionality reduction 	<ul style="list-style-type: none"> • Bias if model is incorrect
Decision tree	<ul style="list-style-type: none"> • Easily understood by human • Capable of representing the most complex data • Much faster in the training phase 	<ul style="list-style-type: none"> • Not flexible at modelling complex features
Random forest	<ul style="list-style-type: none"> • Easy to use • Fast performance • Robust 	<ul style="list-style-type: none"> • Difficult to interpret the prediction
Support Vector Machine	<ul style="list-style-type: none"> • Easy for implementation • Trade-off between classifier complexity and error can be controlled explicitly • Non-traditional data like strings 	<ul style="list-style-type: none"> • Uncalibrated class membership probabilities • Parameters of a solved model are difficult to interpret

	and trees can be used as input to SVM, instead of feature vectors	
Neural network	<ul style="list-style-type: none"> • Able to handle problems with many features • Ability to classify objects even for very complex features 	<ul style="list-style-type: none"> • Slow in the training and the application phase. • Difficulty to determine the network decision making. • Difficulty to determine the important or worthless features
Fuzzy logic	<ul style="list-style-type: none"> • Ability to handle vague data rather than between precise numerical values • Provide a significant level of parametric flexibility 	<ul style="list-style-type: none"> • minmax rule of the membership function is not the way of human reasoning • Lack of learning mechanism
Petri nets	<ul style="list-style-type: none"> • Ability to show precise and graphical representation 	<ul style="list-style-type: none"> • Restricted parameters or values • Inadequate to address the problems of uncertainty, and imprecision in data
Fuzzy Petri nets	<ul style="list-style-type: none"> • Greater representation ability • Ability to reason using uncertain and ambiguous information • Ability to describe the fuzzy behaviour system • The transparent modelling 	<ul style="list-style-type: none"> • Limited to modelling certain kind of problem • Lack of learning mechanism

Statistical, machine learning, fuzzy logic and formal way of classification using Petri Nets and Fuzzy Petri Nets have been widely used in many applications. For statistical methods, the main advantage is the ease of use in terms of the implementation using simple algorithm. However, it is sensitive to the outliers or irrelevant parameters which affect the performance. For machine learning, it makes predictions of models using black-box methods, which provide less information on the prediction. Therefore, it is difficult to interpret the prediction

and determine the important and worthless features for the prediction. For fuzzy logic, the main advantage is its ability to mimic human reasoning. On the other hand, Petri nets (PN) has the ability to provide a precise and graphical representation in ambiguity data. PNs is one of several mathematical representation for the discreet distributed system (Hoheisel & Alt, 2007). PNs are proved to be quite effective tool for graphical modelling, mathematical modelling, simulation, and real time control with the use of places and transitions. In PNs, the processing of tokens and transitions are inherently Boolean. However, in the FPN, it is generalized to involve continuous variables. This extension makes the nets to be fully in rapport with the panoply of the real-world classification problems (Chen et al., 2002). Fuzzy logic, PN and FPN have similar disadvantages that is the lack in learning mechanism.

2.6 Available Classification Methods in Discriminating Impaired Speech Intelligibility

This section presents the existing intelligibility classification system for impaired speech. Kim et al. (2015), extracts the abnormal variation in the prosodic, voice quality and pronunciation aspects in impaired speech. The performance was evaluated on two speech corpuses which are the NKI CCRT Speech Corpus and the TORGO database. They proposed smoothed posterior score fusion of subsystems which gives the best classification performance, 73.5% for unweighted, and 72.8% for weighted, average recalls of the binary classes.

Khan et al. (2003) proposed SVM using n-fold cross validation in classifying intelligibility of Parkinsonian speakers. The classification accuracy of SVM was 85% in 3 levels of UPDRS-S scale and 92% in 2 levels with the average area under the ROC (receiver operating characteristic) curves of around 91%. Fook et al. (2013) performed classification of the prolongations and repetitions among speakers with stuttering. They reported that SVM gives

the best classification accuracy of 95% using the LPC, MLFF and PLP features. From the existing literature, there is no single attempt in experimenting FPN for classifying impaired speech features. Table 2.7 summarizes the available classification system in discriminating speech intelligibility.

University of Malaya

Table 2.7: Summary of available speech intelligibility classification for impaired speech

Research	Data	Stimuli	Features	Classifier	Result
Kim et al. (2015)	<ul style="list-style-type: none"> • NKI CCRT - 17 sentences spoken by 55 speakers • TORGO - 162 sentences of Grandfather passage - 460 sentences from the MOCHA 	Sentence	<ul style="list-style-type: none"> • Prosodic features – pitch contours • Spectral feature – formant, MFCC, phone duration • Perturbation features – jitter, shimmer and harmonics to noise ratio (HNR) 	<ul style="list-style-type: none"> • Support Vector Machine (SVM) • Random Forest • Linear Discriminant Analysis (LDA) classifier • <i>k</i>-nearest neighbour (KNN) 	Best classification accuracy: 73.5% for unweighted, and 72.8% for weighted
Khan et al. (2013)	<ul style="list-style-type: none"> • 240 speech samples • 60 Parkinsonian speakers, 20 control speakers 	Paragraph	<ul style="list-style-type: none"> • Measure of phonatory symptom - Cepstral difference • Measures of articulatory symptoms - MFCC • Measures of prosodic symptoms - F0, Spectral dynamic 	<ul style="list-style-type: none"> • Support Vector Machine (SVM) 	Overall dataset: 83%
Fook et al. (2013)	<ul style="list-style-type: none"> • 77 speech samples of prolongation • 94 speech samples of repetition • 39 stuttered speakers 	Monosyllabic words	<ul style="list-style-type: none"> • MFCC • LPC • PLP 	<ul style="list-style-type: none"> • Support Vector Machine (SVM) • Linear Discriminant Analysis (LDA) classifier • <i>k</i>-nearest neighbour (KNN) 	Best classification accuracy: SVM = 95%

2.7 Summary

This chapter reviews the issues related to the impaired speech characteristics and features, automatic speech detection as well as classification methods in the context of speech intelligibility detection. From the review of the literatures, it can be concluded as follows;

- The reduction of intelligibility in impaired speech is due to several reasons such as imprecise articulation, severity of impaired speakers and speech variability.
- From the aspect of impaired speech database, there are many available databases in the literature, with majority of them for English language. However, there are many positive progress for other languages such as Mandarin, French and Korean. For Malay language, the progress is still in its infancy.
- The speech features play an important role in discriminating speech. This is because, these features correlate to the speeches which carry the meaningful information. Therefore, selecting the relevant speech features is essential. The justification of speech features selection is discussed in Chapter 3.
- There are many classification methods available in the literature. Classification methods are important component in the detection task. Selection of suitable methods is important in improving the detection performance. The justification of selecting the suitable classification methods is discussed in Chapter 3.

CHAPTER 3 RESEARCH METHODOLOGY

This chapter discusses the methodology carried out throughout this research. At the beginning of this chapter, a discussion of findings in LR is presented. Later, the tasks carried out for developing the proposed automatic intelligibility detection of impaired speech are presented as follows;

- Development of speech corpus
- Speech data analysis
- Speech intelligibility measurement
- Development of speech intelligibility detection
- The baseline classification methods
- The proposed FPN classifier
- Evaluation

3.1 Findings of LR

This section discusses the findings from the literature review presented in Chapter 2.

3.1.1 Speech Corpus of Speech Impaired Children

Saz (2011), mentioned that it is strongly necessary to evaluate the real need of building a new corpus for speech research because the process of speech acquisition are time consuming and involves many people, regardless the experimenter or the speakers. The acquisition process becomes more challenging for this research as it involves children with speech impairments due to the severity level, physical and emotional that causes speech variability. Severity of speech impairments not only differs from

individuals; it can also vary for a single speaker depending on the time of day, fatigue, stress or other personal and environmental factors.

One of the problems identified from the literature review is the lack of a good impaired speech corpus for Malay language. Compared to major languages such as English, research in this area for Malay language is still in its infancy and not so favourable due to unavailability of the corpus. The motivation of this work is to fill the gap and provide novel resources to the entire community.

Experimental design in the context of phonetics involves making choices about the speakers, materials, number of repetitions and other issues in such a way that the validity of a hypothesis can be quantified and tested statistically (Harrington, 2010). In speech acquisition process, it is important to consider those issues in order to have a good corpus in terms of quality, which can then lead to better recognition by the ASR system. Several requirements and their rationale have been identified in this process as follows;

Requirements	Purpose/Rationale
Assure low noise and environmental distortions	To make sure the quality of speech samples as noises and environmental distortions can affect the ASR system performance
The speaker need to speak naturally	Speaking styles affects the speech characteristics. Speaking naturally like daily conversation without different speech rate and emotional content to elicit the mispronunciations in a realistic speech
Balance in terms of gender	To be gender independent
Balance in terms of age	To be age independent

Several diagnosis	To cover different types of speech disorders that produce variation in speech and error patterns
Balance in terms of severity level	To elicit the error patterns and mispronunciations that basically occur at different degree of impairments

Further discussion and justification on the development of the speech corpus for Malay speaking children with speech impairments is presented in **Section 3.2**.

3.1.2 Significant Speech Features

A major concern in feature selection is describing the relevance of the features to the problem of interest. The feature selection method attributes the relevance of a subset to the processing task. Despite the belief that a higher number of features provides more discriminating power to the classification, in practice, as a limited amount of training data is accessible, more features slow down the process and classifiers are prone to overfitting. Using irrelevant features degrades the learning performance. One of the key aspects of the feature selection research area is evaluating the advantage of each feature.

In this research, the selection of the speech features is based on the purpose of capturing the abnormal variation in the aspect of prosodic, voice quality and pronunciation of pathological speech as proposed in Kim et al. (2015). Six speech features are chosen for identifying aspects of speech after applying threshold based speech segmentation. These speech features are identified from the literature review. In the aspect of prosodic, three of prosodic features which are F0, energy and zero crossing rate (ZCR) have been chosen. These features are basically used to analyze the variation in harmonic frequencies, which is basically irregular vibration of vocal folds due to unperiodic flow of air through lungs (Butt, 2012).

For voice quality, jitter absolute and shimmer absolute are chosen. Jitter is the frequency perturbation, while shimmer is an amplitude perturbation. Both features are important in voice quality measurement and serves as index of vocal stability (Russell, 2015). Excessive jitter and shimmer cause hoarseness, harsh or rough voice quality. Normal voices are usually less than 1% frequency variability, while, a mean cycle-to-cycle amplitude difference of 0.7 dB or less variation or less than 7% of mean amplitude is normal (Russell, 2015).

Impaired speeches contain higher pronunciation variations that contribute to intelligibility loss. Therefore, pronunciation features for intelligibility classification is considered. For classifying the pronunciation, the most common spectral features, MFCC is chosen. Methods used in the selection of significant speech features are presented in **Section 3.3**.

3.1.3 Speech Intelligibility Measurement

In this research, we use two types of intelligibility measurements which are the subjective rating using human experts and automatic measurement using the automatic speech recognition (ASR). Subjective rating of intelligibility was then compared to the automatic rating from ASR system. A statistical analysis was conducted to determine the significant correlation of intelligibility scores of subjective rating and the automatic rating of ASR system using Pearson correlation test. Further discussion on the selected measurement methods and its justification are discussed in **Section 3.4**.

3.1.4 Speech Classification Method

Based on the investigation and reviews of the related literature in **Chapter 2**, this research proposes to develop the automatic speech intelligibility detection for speech impaired speakers using Fuzzy Petri Nets (FPN) formalism as classifier. An important

advantage is that FPN is its ability to reason using uncertain information and to describe the fuzzy behaviour system with transparent modelling and allows for greater knowledge representation ability.

A good knowledge representation is important in knowledge based system like speech detection to design formalism of complex system which make it easier to design and implement. Knowledge in the context of speech intelligibility is the speech features. For speech intelligibility detection, a knowledge representation scheme based on Fuzzy Petri Net with fuzzy inference algorithms is used. A simple graphical Petri net notation and a well-defined semantics displaying the process of reasoning through inference trees are used for visualization of the knowledge base and explanations of derived conclusion. The knowledge representation formalism has the ability to show a probability of concepts and relations (Ivasic-Kos et al., 2014). It is unambiguously develop a relationship or mapping from speech features to the intelligibility classes. This convenience makes the classification procedures transparent and easily understand as opposed to a black box like most statistical and machine learning.

One of the major problems of degradation in the detection ability is that most patterns belong to the anti-class. It becomes more challenging with limited data or small dictionary for impaired speeches as statistical methods or machine learning need large amount of data for training. FPN uses the rule inference or reasoning which closely resembles to a human mind in making decision, unlike many other method such as statistical and machine learning that depend on data size, where arbitrarily produce the result most likely based on training and input data. It has the capability in extracting inferences and analyzing different rules. Therefore, it is used to adjust the membership functions for each feature vector in a dynamic environment to infer the rules, like in a human mind's decision. This research therefore investigates the potential solutions

available for the development of the automatic speech intelligibility detection for speech impaired speakers using FPN with the ability to classify speeches with high degree of accuracy. Discussion on the selection of the appropriate method and its justification are further discussed in **Section 3.6**.

3.2 Development of Children's Impaired Speech Corpus

This research developed the continuous speech database with dictation speaking style from children with speech impairments.

- Target population (Speaker)

The speaker selection process started by listing potential speakers based on the selection criteria given. Potential children were screened against the selection criteria, which are: (1) aged between 8 years and 12 years; (2) native Malay speaker; (3) gender balance; and (4) able to understand instructions. The intention was to count on a set of children balanced in terms of severity level, diagnosis, gender and age. According to the Convention on the Rights of the Child (CRC) and Malaysia's Child Act 2001, a child essentially means a human being below the age of 18 years unless under the law applicable to the child, majority is attained earlier. The age selection is limited to 12 years old, as most of the boys would reach their puberty after the age of 12 and would include hoarseness in their voice. This condition will affect the acoustic properties and give significant differences across gender. The potential children were assessed with the collaboration of the speech language pathologists (SLPs) and teachers to make sure that they are able to follow through the recording process.

- Speech stimuli

The speech stimuli were provided in Malay language. Malay language belongs to the western subfamily of Malayo Polynesian languages, also known as Austronesian languages (Green & Pawley, 1966). It is used by 500 million people as spoken language, mostly in Malaysia, Indonesia, Brunei, Singapore and southern Thailand (Tan, 2012). Malay language is divided into many dialects. However, this research is focuses on the Standard Malay, which refers to the national norm or prestige dialect, which is also designated as the official language in Malaysia (El-Iman & Don, 2005). Malay is also an official language in Brunei, Indonesia and Singapore. Malay is popularly known as a phonetic language, which means that Malay is actually pronounced very much as it is written in the spelling (Mustafa, 2012).

Table 3.1: Malay consonants

Manner of articulation	Place of articulation							
	Labial			Alveolar			Velar	
	Bilabial	Labio-dental	Dental	Alveolar	Alveo-palatal	Palatal	Velar	Glottal
Plosive	p b			t d			k g	ʔ
Fricative		f v	θ ð	s z	ʃ		x ɣ	h
Affricative				tʃ	dʒ			
Approximant				r				
Lateral				l				
Nasal	m			n	ɲ	ŋ		
Glide	w				j			

Table 3.1 shows the 27 consonant phonemes in Malay according to its manner and place of articulation. Table 3.2 shows the 6 Malay vowel phonemes which are classified according to its height and backness of the tongue.

Table 3.2: Malay vowels

Height	Backness		
	Front	Central	Back
Close	i		u
Mid	e	e'	o
Open		a	

- Speech corpus transcription and labelling

The recorded speech samples were analyzed and labelled for the mispronunciation manually by three expert transcribers who have more than three years' experience in speech sciences. The recorded speeches are evaluated by listening to the audio files for validating the quality of the recorded speech signal. This is also required to evaluate the speech samples and to study the variation of pronunciation in impaired speeches. Table 3.3 shows the transcribers' profile.

Table 3.3: The speech transcribers' profile

Transcriber	Age	Experience (years)
T1	24	4
T2	24	4
T3	24	4

The transcribers also identify the phonological processes to examine the detailed mispronunciations in impaired speech in terms of deletion, substitution, assimilation and others. Phonological processes involve patterns of sound errors that a child would use to simplify speech as they start learning to talk. However, phonological processes in children with speech impairment causes the following;

- The excessive use of phonological processes
- Multiple phonological processes are exhibited together
- Increases the child’s unintelligibility making them difficult to understand

Table 3.4 shows the common phonological processes which describe the systematic changes that affect entire phoneme classes or phoneme sequences.

Table 3.4: Phonological processes to describe phoneme changes (Hodson (1980); Ingram (1981); Shribert & Kwiakowski (1981)) Kahn (1982)

DELETIONS	
1. Initial consonant deletion	at/hat
2. Final consonant deletion	no/noze
3. Consonant cluster reduction	tap/stop
SUBSTITUTIONS	
1. Stopping	ton/sun
2. Voicing/devoicing	die/tie
3. Gliding	ju/shoe
4. Fronting/backing	dum/gum
5. Affrication/deaffrication	chew/shoe
ASSIMILATION	
1. Progressive	beb/bed
2. Regressive	lellow/yellow
3. Velar assimilation	gog/dog
4. Labial assimilation	beb/bed
5. Alveolar assimilation	lellow/yellow
6. Nasal assimilation	neon/pencil
OTHER (infrequent)	
1. Vocalization	bado/bottle
2. Weak syllable deletion	asks/ask
3. Transposition	mud/mother

4. Vowel naturalization	op/stop
5. CC deletion	wawa/water

The procedures of developing speech corpus are further presented in **Chapter 4**.

3.3 Selection of Significant Impaired Speech Features

The measurement of relevant speech features of production in speech impairments has been reported in the literature (Wertzner et al., 2005; Davis, 1978). Wertzner et al., 2005 reported that the most important vocal speech features for clinical use are the measurement of vocal extension profile; frequencies and intensity, noise, acoustic spectrograph; fundamental and formant frequencies and perturbation index; jitter and shimmer. Therefore, this analysis investigates the acoustic features in relation to intelligibility deficits; that include formant frequencies, intensity, fundamental frequency (F0) and perturbation features (jitter, shimmer) in children with speech impairments.

Three methods were used in the data analysis as follows;

- Acoustic analysis
- ASR performance
- Statistical analysis

In acoustic analysis, the analysis involves speech features such as; F1, F2 (Hz), F0 (Hz), intensity (dB), jitter (%) and shimmer (%). The analysis was performed with six Malay vowels /a/, /e/, /i/, /o/, /u/ and /ə/ extracted from selected short sentences. The selected words are disyllabic, which means that the final syllable is pronounced with

prolonged vowel duration. Each vowel sound was segmented into 150 milliseconds. The acoustic analysis was performed using the Windows-based version of Praat software (Boersma & Weenik, University of Amsterdam, The Netherlands). The statistical analysis was performed using the Windows-based SPSS 12.0. Statistical analysis was conducted to determine the significant group mean differences of F1, F2, F0, intensity, jitter and shimmer between the children with and without speech impairments using the ANOVA test.

The analysis are further discussed in **Chapter 4**.

3.4 Subjective and Automatic Measurement of Speech Intelligibility

Subjective measurement is a common measure of speech intelligibility in clinical domain. For the subjective measurement, we apply the same format from the guidelines for assessing speech by The Tennessee Department of Education (*Appendix B: Speech Intelligibility Assessment Form*). It was performed by a panel of Speech Language Pathology (SLP), which is a human expert in speech and hearing language. They are chosen to assess the speech intelligibility due to their expertise in judging the impaired speech. However, due to some limitations in assessing speech by SLP such as high cost and time consuming, there was effort to use automatic assessment using ASR system as proposed in (Shuster et al., 2005). Apart from the subjective measurement, this research performed the automatic way of assessing speech intelligibility using ASR system.

For automatic speech measurement, state of the art ASR system is developed. For clinical use, the speech recognition system will be adapted to disturbed voices and can also be applied to other languages. The motivations for the use of ASR system in measuring the speech intelligibility are as follows;

- The observation that word recognition performed by ASR system can be thought of as machine intelligibility (Liu et al., 2006)
- The positive findings that suggest good correlation between human intelligibility and machine recognition (Chernick et al., 1999; Jiang et al., 2002; and Liu et al., 2006)

The state of the art ASR system for impaired speech that was built is a speaker independent ASR system based on HTK toolkit (CUED, 2009). The HTK toolkit is the most widely used toolkit for developing HMM based ASR system that was originally developed by Steve Young in 1989 at the Speech Vision and Robotics Group of the Cambridge University Engineering Department (CUED). HTK toolkit supports different speech data formats such as speech recognition technology and feature extraction techniques. The reason why HMM is a preferred method in ASR system are as follows (Aarnio, 1999);

1. HMM has strong practical algorithm and mathematical basis for training and recognition.
2. HMM has the ability to handle those conditions and does not need many assumptions for uncertainties condition.
3. HMM is capable to reduce the degradation in performance when moving from speaker-dependent to speaker-independent (Neto et al., 1995).
4. HMM has the ability to match speech production variability (Ynoguti et al., 1998).
5. HMM has the ability to present the speech signal's time sequential order for different speakers (Mao et al., 2007).

The ASR framework is illustrated in Figure 3.1. The speech independent acoustic model is developed by training the unimpaired speeches, the Control Speech (CG) database. The speech recognizer is tested using the impaired speeches, the Speech Impaired (SIG) database. In ASR system speech training, the HMM is the most preferred method for the development of speech acoustic model for ASR systems (Juang & Rabiner, 2005). The recognition result presents here is measured using the Word Error Rate (WER) and Word Recognition Accuracy (WRA).

The method of measuring speech intelligibility using subjective and automatic approach are further discussed in **Chapter 4**.

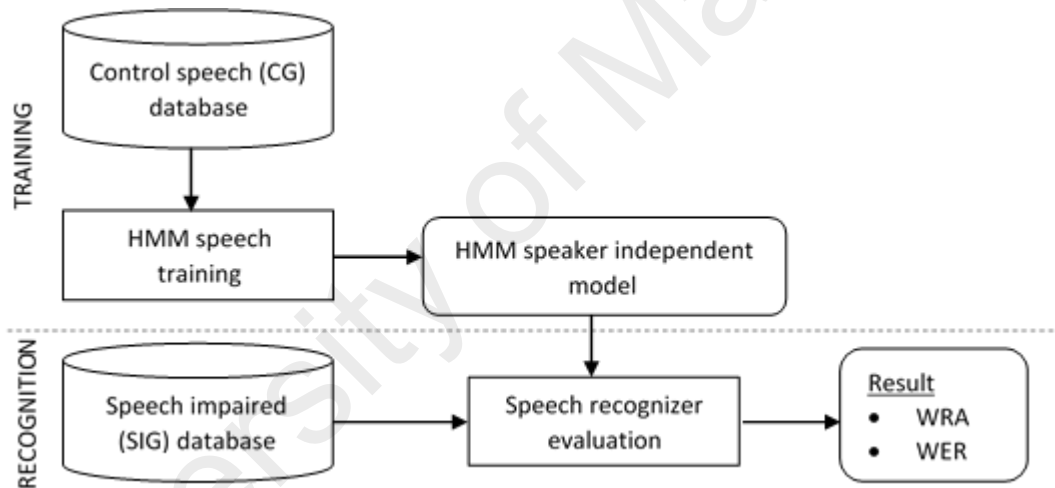


Figure 3.1: The state of the art ASR system framework

3.5 Automatic Speech Intelligibility Detector

In this research, a standard automatic speech intelligibility detector is developed. As discussed in Section 2.1.2, it is important to extract knowledge base front-end of Detection Based Automatic Speech Recognition (DBASR) in improving the classification performance (Li et al., 2005). The relevant knowledge based for impaired

speech is the significant speech features that carries meaningful information of abnormal speech pattern. This section presents the framework of the Automatic Speech Intelligibility Detection. Figure 3.2 shows the overall framework of the system. Basically, the feature extractor extracts speech signals and produce speech features. The speech features are then used by the speech feature classifier. It uses speech signal features as input and gives phonological attribute probability presence as output. The speech features are analyzed by a bank of classifiers, each producing a knowledge scores pertaining to the speech features.

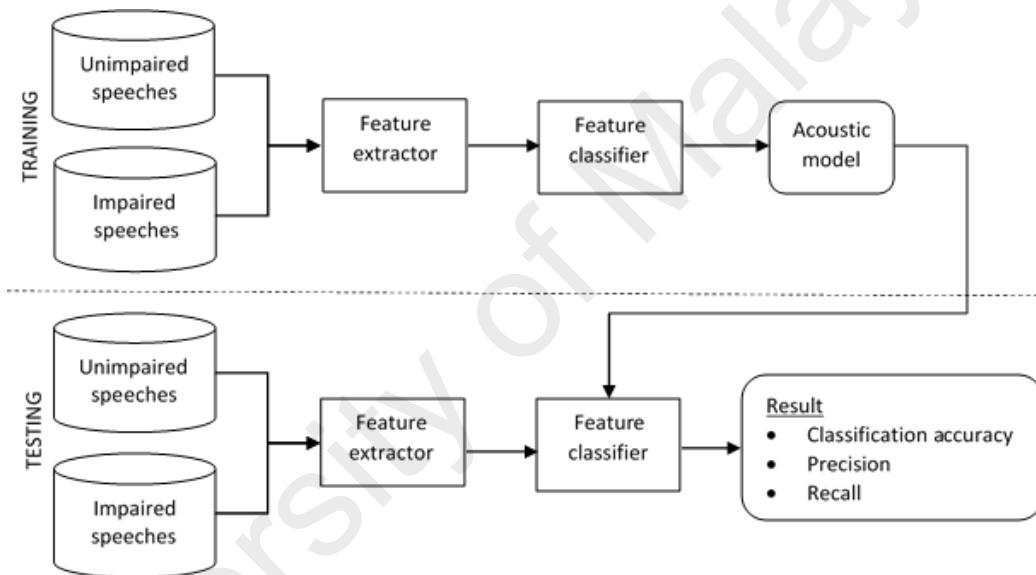


Figure 3.2: Detection system framework

3.5.1 Speech feature extraction

Speech features are extracted using openSMILE toolkit, a feature extraction tool to extract large audio feature spaces in real-time. It is written in C++ and available as both standalone command line executable and dynamic library (Eyben et al., 2015). A major advantage of openSMILE toolkits is that it comes with several reference and baseline feature sets which were used for the INTERSPEECH Challenges (2009-2014) on Emotion, Paralinguistics and Speaker States and Traits, as well as the Audio-Visual

Emotion Challenges (AVEC) from 2011-2013 (Eyben et al., 2015). OpenSMILE has the capability of on-line incremental processing and its modularity. Feature extractor components can be freely interconnected to create new and custom features, all via a simple configuration file. New components can be added to openSMILE via an easy binary plugin interface and a comprehensive API (Eyben et al., 2015).

3.5.2 Selection of Detection Strategies

According to Canterla (2012), there are two ways to justify this detection structure which are as follows;

- It is a special case of segment-based detectors strategy
- The chosen method for the design of sub-word detectors is an adaptation of the standard ASR framework for a two-class problem.

In this research, the proposed detector structure is the segmentation. We have focused on detectors for phonemes and speech intelligibility classes, which are *intelligible* or *not intelligible*. Intelligibility detectors are trained with a database transcribed at the phone level. The anti-class models were built with a combination of all the other phone, for example those that belong to the anti-class. These class and anti-class models are used by a classifier to generate output segmentations.

The further details and development of the automatic speech intelligibility detection are presented in **Chapter 5**.

3.5.3 The Baseline Classification Methods

In this research, four baseline classifiers are chosen which are SVM, RF, LDA and KNN, tested using Matlab. These classifiers are chosen as baseline classifier because they have been tested in existing literature for detecting the impaired speech

intelligibility as discussed in **Section 2.6**. Therefore, it is intriguing to examine the performance of these classifiers using our speech corpus. Further details of the experiments and performance of the selected baseline classifiers are discussed in **Chapter 5**.

3.6 The proposed FPN as a Classification Method in Automatic Speech Intelligibility Detection

The proposed FPN is defined as follows (Chen et al., 2002):

$$\text{FPN} = (P, T, D, I, O, f, \alpha, \beta)$$

where $P = \{p_1, p_2, \dots, p_n\}$ was a finite set of places.

$T = \{t_1, t_2, \dots, t_m\}$ was a finite set of transitions.

$D = \{d_1, d_2, \dots, d_n\}$ was a finite set of propositions.

$$P \ll T \ll D = \Phi, |P| = |D|$$

I and O were the function of set of input and output places of transitions, where

$I: P \rightarrow T$ was the input function, a mapping from transitions to bags of places.

$O: T \rightarrow P$ was the output function, a mapping from transitions to bags of places.

$f: T \rightarrow [0,1]$ was an association function, a mapping from transitions to real values between zero and one.

$\alpha: P \rightarrow [0,1]$ was an association function, a mapping from places to real values between zero and one.

$\beta: P \rightarrow D$ was an association function, a bijective mapping from places to propositions.

A new classification approach to speech intelligibility detection is proposed using fuzzy-Petri-net reasoning generated solution. Reasoning is performed by a fuzzy-Petri-

net detector employing a fuzzy-rule production system design and a fuzzy Petri-net reasoning algorithm, which is developed and implemented in Matlab 2013b. Several stages involves as following;

1. Creating Fuzzy Inference System (FIS), and
2. The proposed FPN classification

This stages are discussed further in the next sections.

3.6.1 Creating Fuzzy Inference System (FIS) using Subtractive Clustering

The purpose of the inference system is to seek information and relationships from the knowledge base and to provide answers, predictions, and suggestions in the way a human expert would provide. The inference engine must find the right facts, interpretations, and rules and then assemble them correctly. This research adopted the Subtractive Clustering method to cluster the dataset. Fuzzy rules are derive from data clustering which is a process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible (Elena, 2013). Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed (Elena, 2013). The main advantage of subtractive clustering is the speed and one pass algorithm to estimate the number of clusters and cluster centres in a set of data (Chiu, 1994).

Types of fuzzy inference system

There are two types of fuzzy inference systems, which are Mamdani and Sugeno.

- Mamdani

Mamdani's fuzzy inference method was proposed in 1975 by Ebrahim Mamdani as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators (mathwork). In Mamdani, the fuzzy implication is modelled by Mamdani's minimum operator, the conjunction operator is min, the t-norm from compositional rule is min and for the aggregation of the rules the max operator is used. To explain the working with this model of FLC, let's consider the example from Rakic (2010) where a simple two-input one-output problem that includes three rules is examined:

Rule1: If x is $A3$ or y is $B1$ then z is $C1$

Rule2: If x is $A2$ and y is $B2$ then z is $C2$

Rule3: If x is $A1$ then z is $C3$

- Sugeno Type

The Sugeno fuzzy inference system is proposed by Takagi, Sugeno and Kang in an effort to develop a systematic approach to generate fuzzy rules from a given input output data set. A typical fuzzy rule in the Sugeno fuzzy system has the form:

If x is A and y is B then $z = f(x,y)$

Where A and B are fuzzy sets in the premise, and $z = f(x,y)$ is a consequent. Usually, $f(x,y)$ is a polynomial in the input variables x and y , but it can be any function as long as it can appropriately describe the output of the system within the fuzzy region specified by the premises of the rule. When $f(x,y)$ is a first order polynomial, the resulting fuzzy inference system is called a first order Sugeno fuzzy model .

Table 3.5 summarizes the strengths of both Mamdani and Sugeno (Reyes, 2012). Basically, Sugeno FIS is similar to the Mamdani method in many areas. The first two

parts of the fuzzy inference process, fuzzifying the inputs and applying the fuzzy operator, are exactly the same. The main difference between Mamdani and Sugeno is that the latter's Sugeno output membership functions are either linear or constant.

Table 3.5: Differences between Mamdani and Sugeno type FIS (Reyes, 2012)

Method	Advantages
Mamdani	It is intuitive It has widespread acceptance It is well suited to human input
Sugeno	It is computationally efficient It can be used to model any inference system in which the output membership functions are either linear or constant It works well with linear techniques It works well with optimization and adaptive techniques It has guaranteed continuity of the output surface It is well suited to mathematical analysis

Membership function

Fuzzy membership function determine the membership functions of objects to fuzzy set of all variables. A membership function, $\mu_F(x)$ provides a measure of the degree of similarity of an element to a fuzzy set. It is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. There are different shapes of membership function such as triangular, trapezoidal and Gaussian.

- Triangular membership function

Let p, q and r represent the three vertices of the X coordinates and $\mu_A(x)$ represents the Y coordinate in fuzzy set A, where A is the membership value. In Equation 3.1, p: lower boundary and r: upper boundary where membership degree is zero, q: the centre where membership degree is 1.

$$\mu_A(x) = \begin{cases} 0 & \text{if } x \leq p \\ \frac{x-p}{q-p} & \text{if } p \leq x \leq q \\ \frac{r-x}{r-q} & \text{if } q \leq x \leq r \\ 0 & \text{if } x \geq r \end{cases} \quad 3.1$$

- Trapezoidal Membership function

The trapezoidal curve is a function of μ_A of vector x , and depends on four scalar parameters p , q , r and s where p and s allocate the "feet" of the trapezoid and parameters q and r allocate the "shoulders." As shown in Equation 3.2.

$$\mu_A(x: p, q, r, s) = \begin{cases} 0 & \text{if } x \leq p \\ \frac{x-p}{q-p} & \text{if } p \leq x \leq q \\ \frac{r-x}{r-q} & \text{if } q \leq x \leq r \\ 0 & \text{if } x \geq r \end{cases} \quad 3.2$$

- Gaussian Membership function

The Gaussian curve is a function of μ_A of vector x , and depends on three scalar parameters p , q and s , where p : center and q : width and s : fuzzification factor (in expression $s=2$). The gaussian membership function μ_A of vector x have been represented by Equation 3.3.

$$\mu(x: p, q, s) = \exp \left[-\frac{1}{2} \left| \frac{x-p}{q} \right|^s \right] \quad 3.3$$

Membership functions were chosen by user's arbitrarily in the past, normally based on the user's experience. Now, membership functions are commonly designed using optimization procedures. The number of membership functions improves the resolution at the cost of greater computational complexity. They normally overlap in expressing

the degree of membership of a value to different attributes. In speech research, the Gaussian type membership function have been widely used (Nereveetil et al., 2014; Zeng et al., 2008; Culebras et al., 2006; Edirisinghe & Sonnadara, 2005).

In this research, the proposed fuzzy inference system is the Sugeno using *anfis*, which is the training routine for Sugeno type FIS with the default membership function is Gaussian. *Anfis* uses a hybrid learning algorithm to tune the parameters of a Sugeno-type FIS. The algorithm uses a combination of the least-squares and back-propagation gradient descent methods to model a training data set. *Anfis* also validate models using a checking data set to test for overfitting of the training data (Mathworks).

3.6.2 The Proposed FPN Classification

In the proposed FPN classification, there are several tasks involved. First, the modelling of Petri Nets in relations to speech features, fuzzy rules and intelligibility classes. Second, the classification of speech intelligibility uses FPN as classification method. In realizing the process of FPN in classifying the speech intelligibility, the selected tools and its usage are described as follows;

1. ***PN Editor*** - To model the Petri Nets for intelligibility detection. The Petri Nets model is exported to the Petri Net Markup Language (PNML) file.
2. ***PNML-2-GPenSIM converter*** - To convert Petri Nets model in .pnml to GPenSIM files which are the MSF, PDF and TDF.
3. ***GPenSIM*** - To communicate with the FIS engine and run the simulation of FPN.

This approach helps to organize a work in a modular way, to use standard libraries and to build own tools. In other words, one is no longer using a 'universal' tool but

he/she is programming his/her own tool with support in a modelling and visualization stage. This is more convenient because no tool is universal enough.

There are a wide range of tools for modelling and implementing FPN that are available with accessible source code. However, there some of tools that are relatively large, difficult to modify and platform dependent. The main advantage of the three selected tools are the compatibility to the Matlab development environment. Matlab is chosen due to the following advantages:

- Matlab runs on many platforms such as Windows, Unix and Linux
- It is quite easy to implement and run algorithm even for beginners.

Further discussion on the process with the selected tools are discussed in the next sections.

Petri Nets modelling using PN Editor

In this research, PN Editor tool is used to model the Petri Nets. PN Editor is developed as graphical interface for Matlab toolbox for Petri nets, which allows to draw discrete PNs, continuous Petri nets, hybrid Petri nets and extended hybrid Petri nets (Svadova et al., 2004). PN Editor is JAVA based application which is designed to be platform independent.

PN Editor supports Petri Net Markup Language (PNML). PNML is XML-based interchange format for Petri nets (it determines Petri nets saving format) and it is described in (Svadova & Hanzalek, 2004). As it is possible to import/export files in PNML format, PN Editor allows maximum preservation of compatibility with other tools that supports PNML.

The editor was originally designed as an environment for graphical interpretation of Petri nets, which is transformed to matrix form suitable for processing in Matlab. Since the use of some functions in Matlab are limited (e.g. displaying possibilities of Matlab), PN Editor can be extended by plugins up to specific user requirements.

PNML-2-GPenSIM converter

The PNML-2-GPenSIM converter is built on MATLAB platform. MATLAB offers a set of functions for reading and interpreting XML files, starting with the function 'xmlread' that reads an XML document and returns a Document Object Model (DOM) node. From the DOM node, the elements of the node (such as 'place', 'transition' and 'arc') can be visited recursively, extracting the names of the elements, the initial marking (in case of place element), the source and the target (in case of an arc element). The following steps are involved in the PNML-2-GPenSIM conversion:

1. Convert XML file to MATLAB structure and get the root of the DOM tree
2. From the root tree, recursively visit the child nodes to get the PNML structure
3. From the PNML structure, get the 'net' child and start extracting Petri Net structure (places, transitions, and arcs)
4. Write the Petri Net structure into the GPenSIM files MSF and PDF.

GPenSIM

The most important reason for developing FPN using GPenSIM and the most advantage of it is its integration with the MATLAB environment, so that harness diverse toolboxes can be harnessed in the MATLAB environment. Other advantages of GPenSIM are as follows (Davidrajuh, 2012);

- Ease of use: anyone with limited mathematical and programming skills should be able to use the tool
- Flexible: possibility of modelling and simulation of discrete event systems in any domain (whether in engineering, business, or economics)
- Extensible: modeller extend or replace any functionality available in the tool, and also add newer functionality
- Compact simulation code.

In this research, experiment with FPN was performed by writing a user M-file that combines GPenSIM with Fuzzy Logic toolbox. In GPenSIM, 3 M-files created are Main Simulation File (MSF), Petri Net Definition Files (PDFs) and Transition Definition Files (TDFs). Figure 3.4 shows the integration of GPenSIM and Fuzzy Logic toolbox in the MATLAB environment.

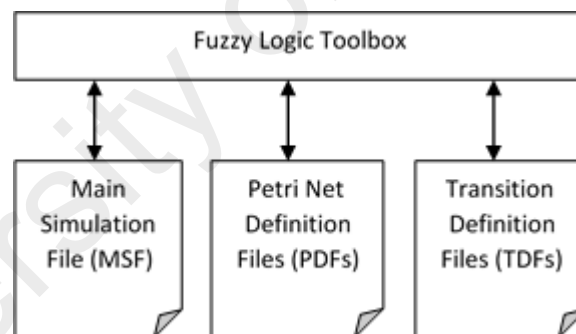


Figure 3.3: Integrating GPenSIM with the Fuzzy Logic Toolbox

The methodology for creating a Petri net model consists of the following three steps:

1. Defining the Petri net graph in a Petri net Definition File (PDF): this is the static part, and consist of three sub-steps:
 - a. Identifying the basic elements of a Petri net graph: the places,
 - b. Identifying the basic elements of a Petri net graph: the transitions,

c. Connecting the elements with arcs

2. In addition to creating the PDF, TDF_PREs for the transitions must be also created. This is because, there are user-defined conditions attached to the transitions.
3. Assigning the dynamics of a Petri net in the Main Simulation File (MSF):
 - a. The initial markings on the places, and possibly
 - b. Assigning the initial dynamics (initial markings and firing times) and running the simulations.

3.7 Evaluation of Classification Method

A classification method involves assigning a set of categories or labels should be assigned to some data, according to some properties of the data. A confusion matrix is commonly used to represent the prediction results of a classifier.

Table 3.6: A confusion matrix of classification results

Actual class	Predicted class		
	C(i j)	Class=Yes	Class=No
Class=Yes		true positives (TP)	false negatives (FN)
Class=No		false positives (FP)	true negatives (TN)

Table 3.6 shows the example of classification results for a binary classification problem, 'Yes' and 'No' which has 2 rows and 2 columns. Across the top is the predicted class labels and down the side are the actual class labels. Each cell contains the number of predictions made by the classifier that fall into that cell. Correct and incorrect predictions are clearly broken down into the two other cells, respectively. The True Positives happens when the classifier correctly predict 'Yes' as 'Yes'. The True

Negatives which ‘Yes’ are classified as ‘No’. False Negatives which are class ‘Yes’ that the classifier marked as ‘No’. We do not have any of those. False Positives are class ‘No’ that the classifier has marked ‘Yes’. This is a useful table that presents both the class distribution in the data and the classifiers predicted class distribution with a breakdown of error types (Brownlee, 2014). In this research, the measurement used to evaluate the classifiers are the classification error rate, accuracy, precision and recall.

Classification error rate

Two types of classification error used are Type I and Type II being generated during classification process. These types of error typically used in binary classification because it generate the clear notion of positive and negative. This research applies a binary classification where it considers whether the speech is intelligible or not intelligible. Type I error is used to measure the False Positive Rate (FPR) for a given class. For a given class, i , FPR measures at what rate the token incorrectly classified as this class (Rosenberg, 2009). The calculation of FPR is in Equation 3.4:

$$\text{FPR of class } i = p(FPi) = \frac{\text{number of false positives}}{\text{number of negative instances}} \quad 3.4$$

Type II error is used to measure the False Negative Rate (FNR) for a given class. . For a given class, i , FNR measures at what rate the token incorrectly classified as another class. The calculation of FNR is in Equation 3.5:

$$\text{FNR of class } i = p(FPi) = \frac{\text{number of false negatives}}{\text{number of positive instances}} \quad 3.5$$

Accuracy

The evaluation of classifier presented here is using the most commonly used measurement which is Accuracy. The calculation is in Equation 3.6:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad 3.6$$

However, accuracy can be misleading. According to Brownlee (2014), sometimes it may be desirable to select a model with a lower accuracy because it has a greater predictive power on a problem. For example, in a problem where there is a large class imbalance, a model can predict the value of the majority class for all predictions and achieve a high classification accuracy, the problem is that this model is not useful in the problem domain. Therefore, two additional common measurements are used; Precision and Recall.

Precision

Precision is the number of True Positives divided by the number of True Positives and False Positives, in Equation 3.7 as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad 3.7$$

It is the number of positive predictions divided by the total number of positive class values predicted. It is also called the Positive Predictive Value (PPV). Precision can be thought as a measure of classifiers exactness. A low precision also indicates a large number of False Positives.

Recall

Recall is the number of True Positives divided by the number of True Positives and the number of False Negatives, in Equation 3.8 as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad 3.8$$

It is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate. Recall can be thought as a measure of classifiers completeness. A low recall indicates many False Negatives.

3.7.1 Accuracy, Precision and Recall

Accuracy may be the most commonly used performance measure in classification. It can be attractive at first because it is intuitive and easy to understand. However, we should not rely on it too much because most data sets are far from symmetric. Using accuracy is only good for symmetric data sets where the class distribution is 50/50 and the cost of false positives and false negatives are roughly the same.

Both precision and recall work well if there is an uneven class distribution that is often the case. They both focus on the performance of positives rather than negatives, which is why it is important to correctly assign the “positive” predicate to the value of most interests. The precision measure shows what percentage of positive predictions were correct, whereas recall measures what percentage of positive events were correctly predicted. To put it in a different way: precision is a measure of how good predictions are with regards to false positives, whereas recall is a measure of how good the predictions are with regards to false negatives. Whichever type of errors are more important, this is the one that should receive the most attention.

In this research, there are two types of evaluation performed as follows;

1. First, the speech data are evaluated in terms of misclassification rate, classification accuracy, precision and recall for the overall data. This is to observe the overall performance for each classifiers.

2. Second, the classification accuracy is derived for individual speech feature. This is to understand which speech features are salient in contributing toward classification task.

Further information in terms of implementation of the proof of concept, evaluation and discussion of findings of the proposed FPN classification methods in speech intelligibility detection are presented in **Chapter 6**.

3.8 Summary

This chapter highlights the proposed approach for the development of the automatic speech intelligibility detection using FPN. Table 3.7 summarized all the task, methods and approaches discussed in this chapter as well as mapping to the particular chapters for further details of implementation.

Table 3.7: Summary of tasks, methods, approaches and more information on the implementation

Task	Method	Approach	More information
Development of speech corpus	Speech recording	<ul style="list-style-type: none"> • Continuous speech • Short sentences 	Chapter 4
	Speech intelligibility measurement	<ul style="list-style-type: none"> • Subjective measurement • Automatic measurement 	
Speech data analysis	Acoustic analysis	<ul style="list-style-type: none"> • Formant frequencies, • Intensity, • Fundamental frequency (F0) • Perturbation features (jitter, shimmer) 	Chapter 4
	ASR performance	<ul style="list-style-type: none"> • MFCC 	
	Statistical analysis	Significant group mean differences using	

		ANOVA Pearson correlation	
Development of automatic speech intelligibility detection	Feature extraction	<ul style="list-style-type: none"> • Prosodic aspect – F0, energy, zcr • Voice quality – jitter absolute and shimmer absolute • Pronunciation – MFCC 0th to 12th 	Chapter 5
	Detection strategies	Segmentation detection	
The baseline classification methods	<ul style="list-style-type: none"> • Support Vector Machine • Random forest • Linear discriminant analysis • K nearest neighbour 	<ul style="list-style-type: none"> • Baseline performance • Evaluation <ul style="list-style-type: none"> ○ Classification error rate ○ Accuracy ○ Precision ○ Recall 	Chapter 5
The proposed FPN as classification method	Data clustering	Fuzzy C Means	Chapter 6
	Types of fuzzy inference	Sugeno	
	Membership function	Gaussian	
	FPN modelling	PN Editor PNMLtoGPenSIM converter	
	FPN classification	GPenSIM	
Evaluation	<ol style="list-style-type: none"> 1. Classify for all speech features 2. Classify for individual speech features 	Classification error rate Accuracy Precision Recall	Chapter 6

CHAPTER 4 A CORPUS OF MALAY SPEAKING CHILDREN WITH SPEECH IMPAIRMENTS: DEVELOPMENT AND ANALYSIS

This chapter focuses on the development of the speech corpus for Malay speaking children with speech impairments. The tasks carried out to develop the speech corpus are as follows:

- Speaker characterization
- Preparing speech materials and stimuli
- Setting the recording environment and apparatus
- Recording procedure and design
- Developing reference speech corpus
- Human transcription and labelling of the impaired speech corpus
- Speech intelligibility measurements

4.1 Speaker Characterization

30 speech impaired children were selected to take part in the recording session from special schools and spastic centre in Petaling Jaya, Kuala Lumpur, Malaysia. There were 16 male and 14 female whose age ranges between 8 and 12 years old; with the mean age of 10 years old. These children were diagnosed with different types of speech impairment. A professional SLPs assessed the children and classified the severity of speech impairment. The severity level was measured using the Percentage of Consonant Correct (PCC) from narrow phonetic transcription (Shriberg & Kwiatkowski, 1982a; 1982b). The PCC index is determined by the number of correct consonants in the

speech utterances. PCC was measured with division of the number of correct consonants by the total number of consonants in the sample and multiplying the result by 100, in Equation 4.1 as follows;

$$\frac{\text{correct consonants}}{(\text{correct consonants} + \text{incorrect consonants})} \times 100 \quad 4.1$$

The severity of speech impaired individual is determined using the PCC index as follows:

- more than 90% : normal
- 90% - 85% : mild
- 85% - 65% : mild to moderate
- 65% - 50% : moderate to severe
- less than 50% : severe

4.2 Speech Materials and Stimuli

The speech stimuli were set to 51 short, simple and meaningful sentences that contain two to five words in each sentence. The sentences were selected after discussions and consultations with the SLPs and teachers. These sentences were designed to suit the speakers' reading abilities and word familiarity. The use of short sentences is also due to the fact that most of the speech impaired children also suffered from physical and cognitive impairments. Thus, these children become easily fatigued, hesitant and tense when they had to utter long or complex sentences. The short, simple, and meaningful sentences were used in this study to provide sufficient features for analysing the speech features and pronunciation errors.

Table 4.1: Sentences in the corpus with its IPA and SAMPA transcriptions

Sentences	IPA	SAMPA
Itu gajah	itu gadʒah	ih t uw g aa j aa hh
Telinganya besar	təlɪŋəpə: bəsar	t er l ih ng er ny er ber s aa r
Kucing makan ikan	kutʃeŋ makan ikan	k uw ch ey ng m aa k aa n ih k aa n
Misai kucing panjang	misai kutʃeŋ. pandʒaŋ	m ih s ay k uw ch ey ng p aa n jh aa ng
Burung boleh terbang	burəŋ bələh tərbəŋ	b uw r ow ng b ow l ey hh t er r b aa ng
Leher zirafah panjang	lehe. zirafah pandʒaŋ	l ey hh ey r z ih r aa f aa hh p aa n jh aa ng
Wau terbang tinggi	wau. tərbəŋ tɪŋgi	w aw t er r b aa ng t ih ng g ih
Kuda lari laju	kudə lari ladʒu	k uw d er l aa r ih l aa jh uw
Kuda ada empat kaki	kudə adə əmpat kaki	k uw d er aa d er er m p aa t k aa k ih
Singa adalah raja hutan	siŋə. adələh. radʒə. hutan	s ih ng er aa d er l aa hh r aa jh er hh uw t aa n
Giginya tajam	gigiŋə tadʒam	g ih g ih ny er t aa jh aa m
Pisang berwarna kuning	pisəŋ bərwanə kunəŋ	p ih s aa ng b er r w aa r n er k uw n ey ng
Pisang banyak khasiat	pisəŋ bəŋaʔ. asiət	p ih s aa ng b aa ny aa ? kh aa s i aa t
Buah tembikai bulat	buah təmikai. bulat	b uw w aa hh t er m b ih k ay b uw l aa t
Kulit tembikai hijau	kulet təmikai. hidʒau	k uw l ey t t er m b ih k ay hh ih jh aw
Isinya merah.	isiŋə me:rah	ih s ih ny er m ey r aa hh
Ada epal merah dan hijau	adə epel me:rah ɖan idʒau	aa d er ey p er l m ey r aa hh d aa n hh ih j aw
Nenas rasa masam	nənas rasə masam	n er n aa s r aa s er m aa s aa m
Fifi budak perempuan	fifi. budaʔ pə:əmpuan	f ih f ih b uw d aa ? p er r er m p uw w aa n
Saya khabar baik	sajə kabar bæ:ʔ	s aa y er kh aa b aa r b aa ey k
Cita- cita Fauzi menjadi pensyarah	ʃitə ʃitə pauzi məndʒadi pəŋʃara	ch ih t er ch ih t er f aw z ih m er n jh aa d ih p er n sy aa r aa hh
Ini syakir	ini ʃaker	ih n ih sy aa k ey r
Syakir budak lelaki.	ʃake budaʔ lələki	sy aa k ey r b uw d aa ? l er l aa k ih
Dia main bola di padang	diə maen bəla di. padaŋ	d ih y er m ay n b ow la d ih p aa d aa ng

Azizah dodoi anak	azizah dōdōi ana?	aa z ih z aa hh d ow d oy aa n aa ?
Fatin menyanyi dalam hutan	faten mənəni. dalam. hutan	f aa t ey n m er ny aa ny ih d aa l aa m hh uw t aa n
Ambil tauhu itu	ambel tauhu? itu	aa m b ey l t aw hh uw ih t uw
Ibu siram pokok bunga	ibu siram. pōko? buŋə	ih b uw s ih r aa m p ow k o k b uw ng er
Buah zaitun di buat minyak	buah zaiton dibuat. mija?	b uw w aa hh z ay t ow n d ih b uw w aa t m ih ny aa ?
Saya boboi	s:ajə bōbōi	s aa y er b ow b oy
Umur saya tujuh tahun	umɔɪ saʝə tuʝəh taɦon	uw m ow r s aa y er t uw j ow hh t aa hh ow n
Boboi sedang gosok gigi	bōbōi sədaŋ ɡosə? gigi	b ow b oy s er d aa ng ɡ ow s ow k ɡ ih ɡ ih
Yaya orang melayu	jaja. ɔraŋ mələju	y aa y aa ow r aa ng m er l aa y uw
Umur yaya lapan tahun	umɔɪ jaja lapan taɦon	uw m ow r y aa y aa l aa p aa n t aa hh ow n
Vini orang cina	vini. ɔraŋ. ʝina	v ih v ih ow r aa ng chi h n er
Kuih pau perisa vanila	kueh pau pərisə vanila	k uw w ey hh p aw p er r ih s er v aa n ih l aa
Pau sangat sedap	pau saŋat sədap	p aw s aa ng aa t s er d aa p
Gopal orang india	ɡopal ɔraŋ india	ɡ ow p aa l ow r aa ng ih n d ih y aa
Gopal pakai jam tangan	ɡopal pakai ʝam taŋan	ɡ ow p aa l p aa k ay jh aa m t aa ng aa n
Zip seluar rosak	zip səluar rōsa?	z ih p s er l uh w aa r r ow s aa ?
Hari ini hari khamis	hari ini hari kames:	hh aa r ih ih n ih hh aa r ih kh aa m ey s
Duit syiling emas	duet ʝileŋ əmas	d uw w ey t sy ih l ih ng er m aa s
Khairul pandai naik basikal	keirəl pandai nae? basikal	kh ay r uw l p aa n d ay n ay k b aa s ih k aa l
Basikal ada dua roda	basikal. adə duə rodə	b aa s ih k aa l a d er d uw w er r ow d er
Ikan koi dalam kolam	ikan kōi dalam. kōlam	ih k aa n k oy d aa l aa m k ow l aa m
Ikan kaloi enak di makan	ikan kaloi ena? dimakan	ih k aa n k aa l oy ey n aa ? d ih m aa k aa n
Van di luar rumah	ven diluar. ɣumah	v ey n d ih l uw w aa r r uw m aa hh
Rumah di tepi tasik	rumah. ditəpi. tase?	r uw m aa hh d ih t er p ih t aa s ey k

Pantai sangat cantik	pantai saŋat ʃanteʔ	p aa n t ay s aa ng aa t cha a n t ey k
Mereka berkelah di tepi pantai	mərekə bəkelah ditəpi pantai	m er r ey k er b er k ey l aa hh d ih t er p ih p aa n t ay
Api merah menyala	api me.rah mənalə	aa p ih m ey r aa hh m er ny aa l er

The sentences and their transcription based on the International Phonetic Alphabet (IPA) and the Speech Assessment Methods Phonetic Alphabet (SAMPA) are shown in the Table 4.1. These 51 sentences consist of 120 words, 360 syllables and 696 phonemes.

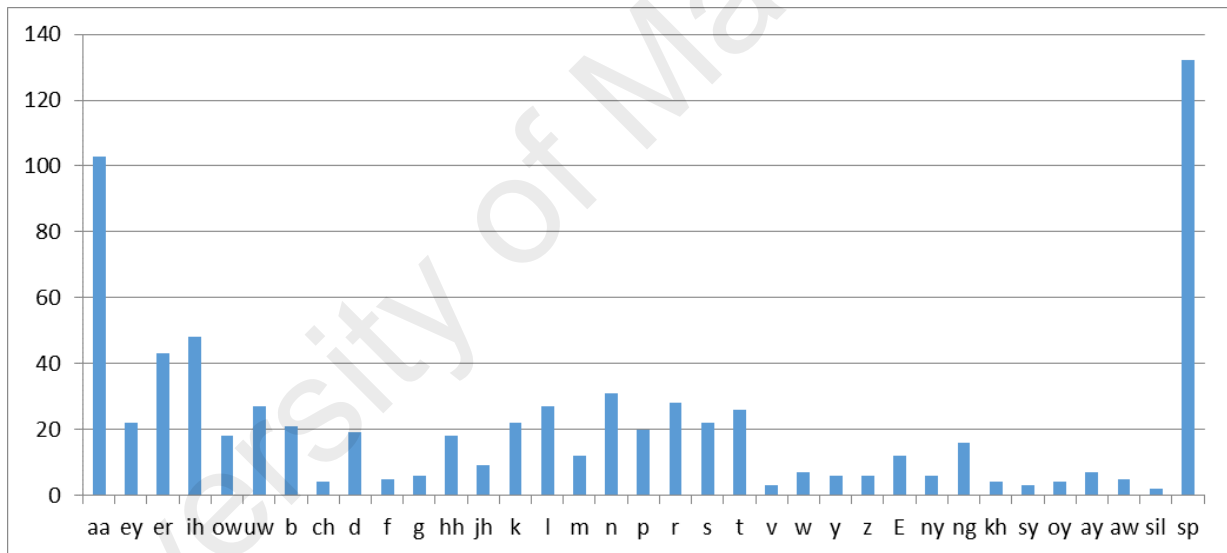


Figure 4.1: Phoneme Distribution in the Speech Samples

To have a good coverage of the speech sounds, the stimuli contain most of the phonemes with the most usual allophones in the context of the target speakers. Figure 4.1 shows the distribution of phoneme in the speech stimuli with reduction of consonant *q* and *x* that are considered unfamiliar among children. The stimuli also covers monosyllabic words to polysyllabic words that consists of vowel (V), vowel-consonant (VC), consonant-vowel (CV) and consonant-vowel-consonant (CVC), clusters of

consonants and diphthongs. Phoneme 'a' is observed as the highest repetitions in the stimuli while phoneme 'v' and 'sy' are the lowest number of repetition. Figure 4.2 shows the example of the speech stimuli that consists of picture and text that were shown to the speakers during recording sessions.



Figure 4.2: Recording interface

4.3 Recording Environment and Apparatus

The recording session took place in a quiet room with a portable sound booth that has a stand microphone for children to speak. The stimuli were presented on a 17-inch Laptop screen. All speech materials were digitized from the audio playback using a 22 kHz sampling rate at 16-bit sample resolution. The lingWAVES Voice Clinic Suite was used to record the speech. The stand microphone is preferred as the speakers might be uncomfortable with a headset microphone. External hard disk is used as a backup storage.

4.4 Recording Procedure and Design

The recording sessions were carried out by placing speakers in the recording room. All speakers were recorded individually, seated at a desk in front of the sound booth. The lingWAVES stand microphone was positioned approximately 4 to 6 inches from

the speaker's mouth. The speech stimuli were displayed to the speakers using a laptop screen. The experimenter was seated beside the speakers to assist in the reading. The session was designed to be fulfilled by the speakers in the corpus with simple meaningful sentences. Each speaker was asked to utter 51 sentences in three repetitions. Each repetition is considered as a session. Therefore, 3 sessions were designed for each speakers and each session was recorded in a different day or big gap of time to reflect intra-speaker variability and avoid the speakers from fatigue and fluctuate emotional state due to long recording sessions.

The sentences were pronounced by the experimenter followed by the speaker. They were encouraged to speak naturally and clearly. As a result, the total amount of impaired speech samples acquired during the whole process was 4,590 utterances in 3.8 hours of recordings including silence. Table 4.2 summarize the impaired speech corpus that has been developed.

Table 4.2: Descriptions of the impaired speech corpus

Speakers' age (years old)	8	6 speakers
	9	6 speakers
	10	6 speakers
	11	6 speakers
	12	6 speakers
Speakers' gender	Male	16 speakers
	Female	14 speakers
Speakers' diagnosis	Cerebral Palsy (CP)	12 speakers
	Hearing impaired	18 speakers
Speakers' severity level	Mild	8
	Mild-moderate	9

	Moderate-severe	6
	Severe	7

4.5 Reference Speech Corpus

From the literature, there is no existing corpus for regular children which can be deemed as an appropriate reference corpus in this research. Ting et al. (2012) studied the acoustic characteristics of regular Malay children ranging from 7 to 12 years old involving 360 speakers. However the data only consists of sustained vowel. This is not appropriate since this research focused at sentences level. The age differences also need to be taken into consideration. The age mismatch could cover effects of the speech impairments because children's voices were always expected to obtain worse performance than adult voices in ASR (Saz, 2012). Therefore, it is necessary to develop our own reference corpus.

Our reference speech corpus consists of speech from 50 unimpaired children (25 males, 25 females) with age ranging from 8 to 12 years old. The recording environment, procedures and speech stimuli were the same with the impaired corpus. The intention was to have a group of speakers that are balanced in terms of age and gender. The selected children were assessed by their teachers to ensure that they are good in literacy. The speakers involved are shown in the Table 4.3 below.

Table 4.3: Number of male and female unimpaired speakers by age

Age	Male	Female	Age	Male	Female	Age	Male	Female
8	5	5	10	5	5	12	5	5
9	5	5	11	5	5			

Each speaker uttered the same 51 sentences in one session of recording supervised by the experimenter. The total amount of speech samples acquired during the whole process was 7,650 utterances in 2.5 hours of recordings including silence.

4.6 Human Transcription and Labelling of the Impaired Speech Corpus

Figure 4.3 shows the frequency of errors according to its category. Backing has the highest frequency of errors in the speech samples with 149 occurrences, followed by liquid gliding (n=123), insertion (n=115) and so on.

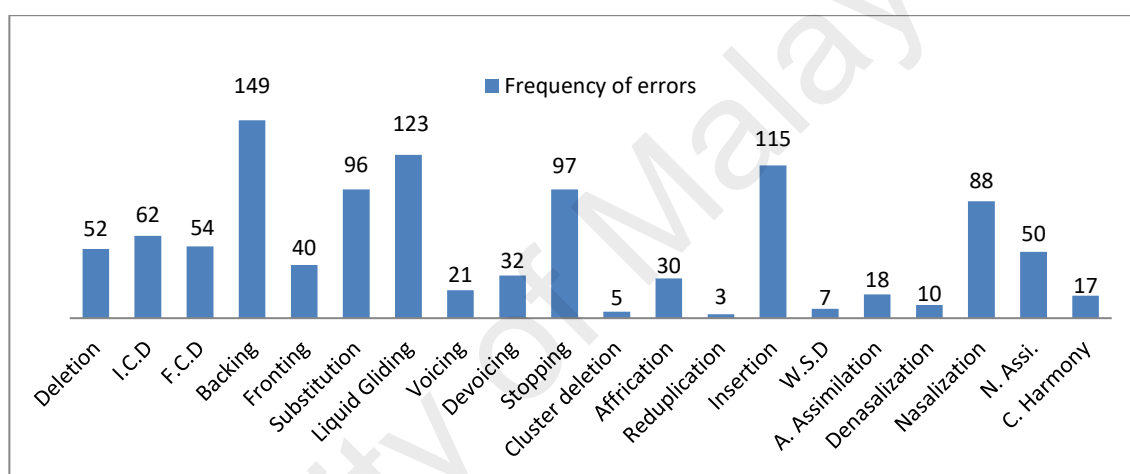


Figure 4.3: Frequency of errors

4.7 Speech Intelligibility Measurement

As discussed in Chapter 3, we have chosen two types of intelligibility measurement measuring the speech intelligibility which are;

- Subjective evaluation by SLPs
- Automatic measurement using ASR system

The following subsections discuss further details in measuring the speech intelligibility using these two types of measurements.

4.7.1 Subjective Evaluation by SLPs

The three SLPs subjectively estimate the intelligibility of the impaired speech of each speaker while listening to a play-back of the speech recordings. The intelligibility assessment form were given to the SLPs to rate the intelligibility of all individual speech samples. There are 30 impaired speakers with 51 short sentences for each speaker. A total of 1,530 speech samples are provided for the subjective evaluation. The average of the scores are obtained based on the consensus among the SLPs. Table 4.4 shows the intelligibility scores given by the SLPs.

Table 4.4: Intelligibility scores by SLPs

Speaker	Intelligibility scores	Speaker	Intelligibility scores	Speaker	Intelligibility scores
SIG01	88	SIG11	86	SIG21	78
SIG02	63	SIG12	79	SIG22	50
SIG03	87	SIG13	68	SIG23	59
SIG04	77	SIG14	42	SIG24	58
SIG05	67	SIG15	56	SIG25	43
SIG06	44	SIG16	87	SIG26	52
SIG07	78	SIG17	33	SIG27	83
SIG08	86	SIG18	86	SIG28	51
SIG09	86	SIG19	20	SIG29	37
SIG10	45	SIG20	87	SIG30	34

4.7.2 Automatic Intelligibility Measurement using ASR

This section discusses the steps in developing speaker independent ASR system in detail, recognizer evaluation as well as the system performance.

4.7.2.1 Data Preparation

The first step in developing speaker independent ASR system is to prepare the data that will be used in the speech training and testing. This section presents the data

preparation which includes tasks such as recording speech data, building task grammar, pronunciation dictionary and transcription files (Young et al., 2006).

```
*/AS1 ITU GAJAH
*/AS2 TELINGA NYA BESAR
*/AS3 KUCING MAKAN IKAN
*/AS4 MISAI KUCING PANJANG
*/AS5 BURUNG BOLEH TERBANG
*/AS6 LEHER ZIRAFAH PANJANG
*/AS7 WAU TERBANG TINGGI
*/AS8 KUDA LARI LAJU
*/AS9 KUDA ADA EMPAT KAKI
*/AS10 SINGA ADALAH RAJA HUTAN
*/AS11 GIGI NYA TAJAM
*/AS12 PISANG BERWARNA KUNING
*/AS13 PISANG BANYAK KHASIAT
*/AS14 BUAH TEMBIKAI BULAT
*/AS15 KULIT TEMBIKAI HIJAU
*/AS16 ISI NYA MERAH
*/AS17 ADA EPAL MERAH DAN HIJAU
*/AS18 NENAS RASA MASAM
*/AS19 FIFI BUDAK PEREMPUAN
*/AS20 SAYA KHABAR BAIK
*/AS21 CITA CITA FAUZI MENJADI PENSYARAH
*/AS22 INI SYAKIR
*/AS23 SYAKIR BUDAK LELAKI
*/AS24 DIA MAIN BOLA DI PADANG
*/AS25 AZIZAH DODOI ANAK
*/AS26 FATIN MENYANYI DALAM HUTAN
*/AS27 AMBIL TAUHU ITU
*/AS28 IBU SIRAM POKOK BUNGA
*/AS29 BUAH ZAITUN DI BUAT MINYAK
*/AS30 SAYA BOBOI
*/AS31 UMUR SAYA TUJUH TAHUN
*/AS32 BOBOI SEDANG GOSOK GIGI
*/AS33 YAYA ORANG MELAYU
*/AS34 UMUR YAYA LAPAN TAHUN
*/AS35 VINI ORANG CINA
*/AS36 KUIH PAU PERISA VANILA
*/AS37 PAU SANGAT SEDAP
*/AS38 GOPAL ORANG INDIA
*/AS39 GOPAL PAKAI JAM TANGAN
*/AS40 ZIP SELUAR ROSAK
*/AS41 HARI INI HARI KHAMIS
*/AS42 DUIT SYILING EMAS
*/AS43 KHAIRUL PANDAI NAIK BASIKAL
*/AS44 BASIKAL ADA DUA RODA
*/AS45 IKAN KOI DALAM KOLAM
*/AS46 IKAN KALOI ENAK DI MAKAN
*/AS47 VAN DI LUAR RUMAH
*/AS48 RUMAH DI TEPI TASIK
*/AS49 PANTAI SANGAT CANTIK
*/AS50 MEREKA BERKELAH DI TEPI PANTAI
*/AS51 API MERAH MENYALA
```

Figure 4.4: The task grammar

Data recording

All the speech data were pre-recorded as discussed in **Section 4.1** until **Section 4.5**.

Task grammar

Task grammar consists of a set of variable definitions followed by a regular expression describing the words to recognize (Young et al., 2006). The words to recognize are the same sentences used in our speech database. 51 sentences were used

for building task grammar for the impaired speakers as well as control speakers as shown in Figure 4.4.

Pronunciation dictionary

Pronunciation dictionary is created by sorting all the words contained in the speech database and also unrelated words for the recognition task to have a phonetically balanced dictionary. The pronunciation dictionary consists of 751 sorted words with their machine readable transcription. Figure 4.5 shows the sample of the pronunciation dictionary.

ABANG	aa b aa ng
ABU	aa b uw
ADA	aa d er
ADALAH	aa d er l aa hh
ADIK	aa d ey E
ADUHAI	aa d uw hh ay
AGAR	aa g aa r
AHAD	aa hh aa d
AHLI	aa hh l ih
AHMAD	aa hh m aa d
AISKRIM	ay s k r ih m
AISYA	ay sy aa
AJAR	aa jh aa r
AKAN	aa k aa n
AKHIR	aa kh ih r
AKIBAT	aa k ih b aa t
AKU	aa k uw
ALAM	aa l aa m
ALI	aa l ih
ALIA	aa l ih y aa
ALISA	aa l ih s aa
ALIYA	aa l ih y aa

Figure 4.5: The sample of pronunciation dictionary

Transcription file

The first step is to generate *Master Label File* (MLF) from prompts file in task grammar. The MLF contain complete transcription of each word. Figure 4.6 shows a sample transcription at word level for the sentence "itu gajah" (that is elephant). Once the word level MLF has been created, phone level MLFs is generated.

```
#!MLF!#
"/AS1.*.*.lab"
ITU
GAJAH
.
```

Figure 4.6: A transcription at word level for the sentence “itu gajah”

To train a set of HMMs, every file of training data must have the associated phone level transcription. Since there is no hand labelled data to bootstrap a set of models, a flat-start scheme was used instead. In order to do this, two sets of phone transcriptions will be required. The set that was initially used will not have short-pause (sp) models between words. Once reasonable phone models have been generated, a sp model will be inserted between the words to take care of any pauses introduced by the speaker. Figure 4.7 (a) shows a sample transcription at phoneme level without a sp model while figure 4.7 (b) shows a transcription at phoneme level with a sp model.

```
#!MLF!#
"/AS1.*.*.lab"
sil
ih
t
uw
g
aa
jh
aa
hh
sil
.
```

(a)

```
#!MLF!#
"/AS1.*.*.lab"
sil
ih
t
uw
sp
g
aa
jh
aa
hh
sp
sil
.
```

(b)

Figure 4.7: A transcription at phoneme level for the sentence “itu gajah” (a) without a sp model (b) with a sp model

4.7.2.2 Feature Extraction

Feature extraction converts the speech waveform to some type of parametric representation that is also known as feature vectors. The parameters used in the MFCC feature extraction are as follows;

```
SOURCEFORMAT = WAV  
TARGETKIND = MFCC_0  
TARGETRATE = 100000.0  
SAVECOMPRESSED = T  
SAVEWITHCRC = T  
WINDOWSIZE = 250000.0  
USEHAMMING = T  
PREEMCOEF = 0.97  
NUMCHANS = 26  
CEPLIFTER = 22  
NUMCEPS = 12
```

The speech WAV file was then separated into frames by multiplication of overlapping Hamming windows. The interval was set to 10 milliseconds (HTK use units of 100ns) and the window length was 25 milliseconds. The first order for pre-emphasis coefficients used in the signal is 0.97 where it should be in the range of $0 \leq k < 1$ (Young et al, 2006). There are 26 channels available in filter bank. The cosine transform was used to reduce the dimensions of the feature vectors from 26 to 12. The speech is parameterized with 12 MFCC and normalized log energy (MFCC=13), plus the delta coefficients (+13) and the acceleration coefficients (+13) yielding a total of 39 components. In this research, the features used are 39-dimensional MFCC. These feature vectors are then used to further process the speech training and recognizing.

4.7.2.3 Speech Training

Figure 4.8 depicts the overall speech training framework. The speech wav files are extracted using feature extraction technique to produce feature vectors. Then, speech training uses the feature vectors, together with HMM prototype, transcription label and pronunciation dictionary to produce HMM model. Transcription label and pronunciation

dictionary have been prepared earlier in data preparation. The derivation of HMM prototype will be presented in this subsection.

In this research, the acoustic model was trained from 186 unimpaired speakers with monophone and triphone HMM models. In monophone acoustic model, we used phoneme HMMs on the unimpaired speech containing 34 monophones with single Gaussian per HMM state. For triphone acoustic model, we use crossword triphone HMMs on the unimpaired speech containing 464 tied states with 12 Gaussian mixtures per state.

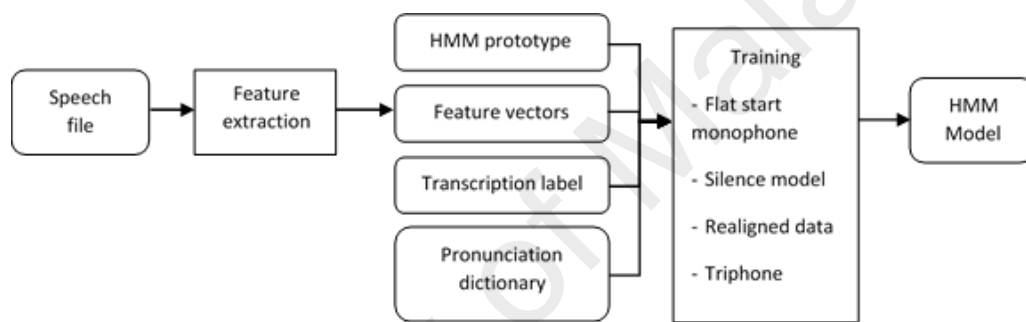


Figure 4.8: Speech training framework

In this section, the detailed steps in speech training are presented. Several steps were conducted such as creating flat start monophones, fixing the silent models, realigning the training data and making triphones from monophones.

Creating flat start monophones

In speech training process, first, we create the prototype of HMM model to define the HMM topology. The topology used is a 3 state left-right where each ellipsed vector is the length of 39 MFCC. All of the mean and variance Gaussian are set to be equal to the global mean and variance of the speech training data. A list of the training files is shown in figure 4.9.


```
./train/mfcc/AS1.FU001.1.mfc
./train/mfcc/AS1.FU001.2.mfc
./train/mfcc/AS1.FU002.1.mfc
./train/mfcc/AS1.FU002.2.mfc
./train/mfcc/AS1.FU003.1.mfc
./train/mfcc/AS1.FU003.2.mfc
./train/mfcc/AS1.FU004.1.mfc
./train/mfcc/AS1.FU004.2.mfc
./train/mfcc/AS1.FU005.1.mfc
./train/mfcc/AS1.FU005.2.mfc
./train/mfcc/AS1.FU006.1.mfc
./train/mfcc/AS1.FU006.2.mfc
./train/mfcc/AS1.FU007.1.mfc
./train/mfcc/AS1.FU007.2.mfc
./train/mfcc/AS1.FU008.1.mfc
./train/mfcc/AS1.FU008.2.mfc
./train/mfcc/AS1.FU009.1.mfc
./train/mfcc/AS1.FU009.2.mfc
```

Figure 4.9: List of the training files

This process is normally used at the initial stage for flat-start training, performed using the HCompVcommand in HTK toolkit. The flat monophones created were re-estimated by refining the parameters of the existing HMMs using Baum-Welch Re-estimation algorithm. Here, we have created 3 state left-to-right HMM for each phone as well as a HMM for the silence model, sil.

Fixing the silence models

To fix the silence models is to add extra transitions from states 2 to 4 and 4 to 2 in the silence model as shown in figure 4.10. Therefore, sp model is added between words in the speech file. The reason is to have a more robust model by allowing individual states to absorb the various impulsive noises in the training data (Young et al., 2006). The backward skip allows this to happen without committing the model to transit to the following word (Young et al., 2006).

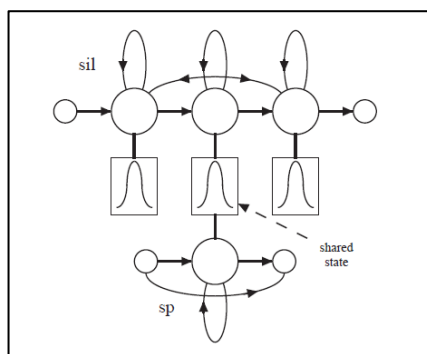


Figure 4.100: Silence model

Realign the training data

The phone models that have been created are used to realign the training data and create new transcriptions with a single invocation of the HTK recognition tool HVite as follows;

```
HVite -l '*' -o SWT -b silence -C config -a -H hmm7/macros \
-H hmm7/hmmdefs -i aligned.mlf -m -t 250.0 -y lab \
-I words.mlf -S train.scp dict monophones1
```

This command uses the HMM models to transform the input word level transcription to the new phone level transcription using the pronunciations stored in the pronunciation dictionary created earlier. The major difference between this operation and the original word-to-phone mapping is creation of transcription process, where the recognizer considers all pronunciations for each word and outputs the pronunciation that best matches the acoustic data (Young et al., 2006).

Making triphones from monophones

Context-dependent triphones can be made by simply cloning monophones and then re-estimate it using triphone transcriptions. The clone command is used to tie all of the transition matrices in each triphone set as follows;

TI T_ah {(*-ah+*,ah+*,*-ah).transP}

TI T_ax {(*-ax+*,ax+*,*-ax).transP}

TI T_ey {(*-ey+*,ey+*,*-ey).transP}

TI T_b {(*-b+*,b+*,*-b).transP}

TI T_ay {(*-ay+*,ay+*,*-ay).transP}

The triphone is described as a-b+c where each model of the form a-b+c in the list, it looks for the monophone b and makes a copy of it. This style of triphone transcription is referred to as word internal. Figure 4.11 shows the conversion from the (a) monophone model becomes (b) triphone models for sentence “itu gajah”.

```
#!MLF!#
"/AS1.*.*.lab"
sil
ih
t
uw
g
aa
jh
aa
hh
sil
.
```

(a)

```
#!MLF!#
"/AS1.FU001.1.lab"
sil
ih+t
ih-t+uw
t-uw
sp
g+aa
g-aa+jh
aa-jh+aa
jh-aa+hh
aa-hh
sp
sil
.
```

(b)

Figure 4.11: Example of conversion from (a) monophone model to (b) triphone model

4.7.2.4 ASR Evaluation

Figure 4.12 shows the framework of speech recognizer evaluation. The speech WAV files are extracted using feature extraction that produces feature vectors. Then, the speech recognizer uses feature vectors, together with the HMM model derived from the training process, the grammar and pronunciation dictionary prepared earlier in data preparation. Later, the recognizer outputs the word from the transcription label that best matches the acoustic data.

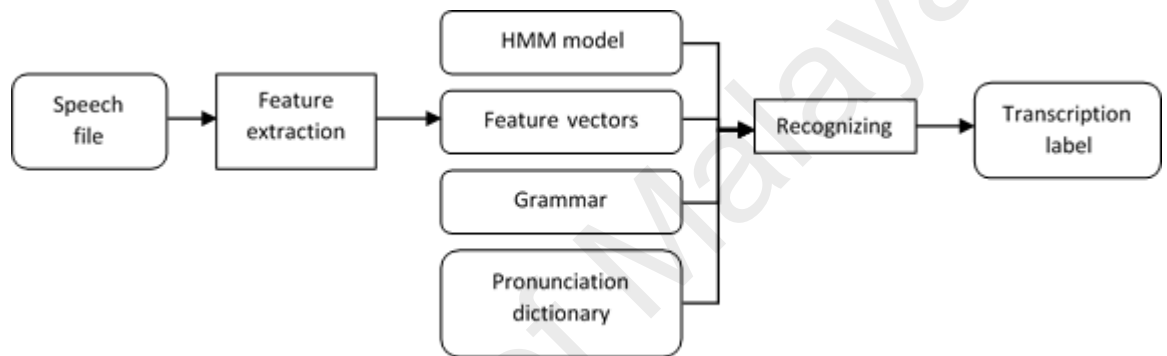


Figure 4.12: Speech recognition framework

In this research, recorded speeches from 30 impaired speakers were used for the recognizer evaluation. The total test data used was 450 utterances. The MLF transcription files for testing was prepared earlier in data preparation, where the procedure is similar to preparing the MLF transcription files for training purpose. The test data were converted into MFCC format as shown in figure 4.13.

```
#!MLF!  
"S1.*.lab"  
ITU  
GAJAH  
.
```

Figure 4.13: Sample of MLF transcription files

4.7.2.5 Result of ASR System

The recognition result presents here is measured using the Word Recognition Accuracy (WRA) and Word Error Rate (WER). Table 4.5 presents the results of speaker independent ASR baseline system.

Table 4.5: Results of ASR baseline system

Speaker	WRA	Substitution	Deletion	Insertion	WER
SIG01	30.61	69.39	0.00	81.63	69.39
SIG02	17.02	78.72	4.26	61.70	82.98
SIG03	67.31	32.69	0.00	76.92	32.69
SIG04	40.38	59.62	0.00	59.62	59.62
SIG05	20.37	79.63	0.00	120.37	79.63
SIG06	11.11	85.19	3.70	96.30	88.89
SIG07	35.19	64.81	0.00	57.41	64.81
SIG08	29.63	68.52	1.85	29.63	70.37
SIG09	29.63	70.37	0.00	46.30	70.37
SIG10	7.41	92.59	0.00	105.56	92.59
SIG11	15.00	85.00	0.00	107.50	85
SIG12	22.22	77.78	0.00	64.81	77.78
SIG13	42.59	57.41	0.00	81.48	57.41
SIG14	18.52	81.48	0.00	96.30	81.48
SIG15	42.31	57.69	0.00	42.31	57.69
SIG16	100.00	0.00	0.00	11.54	0
SIG17	26.92	73.08	0.00	44.23	73.08
SIG18	100.00	0.00	0.00	0.00	0
SIG19	5.77	94.23	0.00	84.62	94.23
SIG20	94.23	5.77	0.00	0.00	5.77
SIG21	31.91	68.09	0.00	97.87	68.09
SIG22	31.48	68.52	0.00	103.70	68.52

SIG23	25.93	74.07	0.00	88.89	74.07
SIG24	24.07	75.93	0.00	101.85	75.93
SIG25	27.78	66.67	5.56	105.56	72.22
SIG26	18.52	81.48	0.00	46.30	81.48
SIG27	20.37	79.63	0.00	46.30	79.63
SIG28	27.78	72.22	0.00	85.19	72.22
SIG29	12.96	83.33	3.70	51.85	87.04
SIG30	12.96	87.04	0.00	98.15	87.04
Average	32.84	66.52	0.64	69.77	67.16

The word recognition accuracy derived from the ASR system is 32.84%. Speakers SIG16 and SIG18 both produce the highest accuracy with 100%, while speaker SIG19 produce the lowest accuracy with 5.77%. The average score for substitution, deletion and insertion are 66.52%, 0.64% and 69.77% respectively.

4.7.3 Discussion: Relationship between Intelligibility Scores and WRA

In Chapter 4, we have determined the intelligibility scores for each of the impaired speaker. In the previous section, we have come up with the speech recognition accuracy produced by ASR. In this section we investigate the correlation between WRA and the intelligibility scores. Based on the analysis shown in Figure 4.14, there is a strong correlation, which the WRA increase with the increment of intelligibility scores. The correlation between intelligibility scores with WRA is significant at $p < 0.05$, $r(0.57) = 0.01$. This indicated a strong correlation between the subjective and automatic evaluation results.

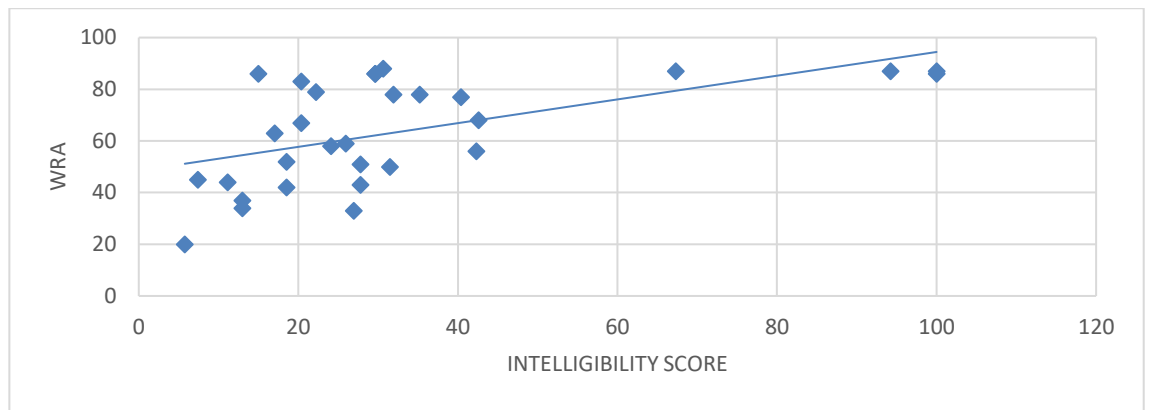


Figure 4.14: The correlations of the WRA with the intelligibility scores

Table 4.6 presents the intelligibility scores differences between subjective and automatic evaluation results. From the table, subjective evaluation performed by SLPs shows higher scores as compared to automatic evaluation by ASR system except for three speakers which are SIG16 (100%), SIG18 (100%) and SIG20 (94.23%). There are increment of 13% for SIG16, 14% for SIG18 and 7.23 for SIG20. Even though there is a strong positive correlation between subjective and automatic measurement, the automatic intelligibility scores given by the ASR system seems to be less reliable for SIG16, SIG18 and SIG20. This is because, the intelligibility score similar or greater than 90% is considered as not impaired. Therefore, the intelligibility scores given by SLPs are considered to be used in the speech corpus.

Table 4.6: Difference of subjective and automatic intelligibility scores

Speakers	Subjective	Automatic	Difference
SIG01	88	30.61	57.39
SIG02	63	17.02	45.98
SIG03	87	67.31	19.69
SIG04	77	40.38	36.62
SIG05	67	20.37	46.63

SIG06	44	11.11	32.89
SIG07	78	35.19	42.81
SIG08	86	29.63	56.37
SIG09	86	29.63	56.37
SIG10	45	7.41	37.59
SIG11	86	15.00	71.00
SIG12	79	22.22	56.78
SIG13	68	42.59	25.41
SIG14	42	18.52	23.48
SIG15	56	42.31	13.69
SIG16	87	100.00	-13.00
SIG17	33	26.92	6.08
SIG18	86	100.00	-14.00
SIG19	20	5.77	14.23
SIG20	87	94.23	-7.23
SIG21	78	31.91	46.09
SIG22	50	31.48	18.52
SIG23	59	25.93	33.07
SIG24	58	24.07	33.93
SIG25	43	27.78	15.22
SIG26	52	18.52	33.48
SIG27	83	20.37	62.63
SIG28	51	27.78	23.22
SIG29	37	12.96	24.04
SIG30	34	12.96	21.04

4.8 Speech Corpus

Table 4.7 shows the details of each speaker with the diagnosis, severity level and intelligibility scores provided.

Table 4.7: Description of the speakers

Speaker	Gender	Age	Diagnosis	Severity level	Intelligibility Scores (Subjective)
SIG01	Female	12	Dysarthria	Mild	88
SIG02	Female	11	Dysarthria	Moderate to severe	63
SIG03	Female	10	Hearing impaired	Mild	87
SIG04	Female	10	Dysarthria	Mild to moderate	77
SIG05	Female	7	Hearing impaired	Mild to moderate	67
SIG06	Female	7	Hearing impaired	Severe	44
SIG07	Female	8	Hearing impaired	Mild to moderate	78
SIG08	Female	10	Hearing impaired	Mild	86
SIG09	Female	11	Hearing impaired	Mild	86
SIG10	Female	11	Hearing impaired	Severe	45
SIG11	Female	13	Dysarthria	Mild	86
SIG12	Female	13	Hearing impaired	Mild to moderate	79
SIG13	Female	12	Hearing impaired	Mild to moderate	68
SIG14	Female	12	Hearing impaired	Severe	42
SIG15	Male	11	Dysarthria	Moderate to severe	56
SIG16	Male	10	Dysarthria	Mild	87
SIG17	Male	12	Hearing impaired	Severe	33
SIG18	Male	8	Hearing impaired	Mild	86
SIG19	Male	11	Dysarthria	Severe	20
SIG20	Male	10	Dysarthria	Mild	87
SIG21	Male	9	Dysarthria	Mild to moderate	78

SIG22	Male	6	Hearing impaired	Severe	50
SIG23	Male	7	Hearing impaired	Moderate to severe	59
SIG24	Male	11	Hearing impaired	Moderate to severe	58
SIG25	Male	11	Hearing impaired	Severe	43
SIG26	Male	12	Hearing impaired	Moderate to severe	52
SIG27	Male	12	Hearing impaired	Mild to moderate	83
SIG28	Male	13	Hearing impaired	Moderate to severe	51
SIG29	Male	13	Hearing impaired	Severe	37
SIG30	Male	10	Hearing impaired	Severe	34

4.9 Analysis of Impaired and Control Speech in Relations to Intelligibility Deficits

There are three types of analysis conducted to observe the differences of speech among SIG and CG group which are;

- Acoustic analysis for formant frequencies, intensity, fundamental frequency (F0) and perturbation features such as jitter, shimmer.
- Word recognition accuracy using MFCC
- Statistical analysis for identifying the significant differences among SIG and CG.

4.9.1 Acoustic Analysis

This section investigates the speech features of impaired speech; that include formant frequencies, intensity, fundamental frequency (F0) and perturbation features such as jitter and shimmer. The selection of these features has been discussed in **Section 3.4**.

The analysis of F1, F2 (Hz), F0 (Hz), intensity (dB), jitter (%), shimmer (%) and HNR (dB) were performed with six Malay vowels /a/, /e/, /i/, /o/, /u/ and /ə/ extracted from the selected short sentences, as shown in Table 4.8.

Table 4.8: Selected sentences and words for vowels extraction

Vowel	Phone	Sentences	Words selected	IPA	Phoneme
/a/	aa	Dia main <i>bola</i> di padang <i>He plays with a ball in the field</i>	Bola (<i>ball</i>)	Bola	b-ow-l-aa
/e/	ey	<i>Leher</i> zirafah panjang <i>The giraffe's neck is long</i>	Leher (<i>neck</i>)	Leher	l-ey-h-ey-r
/i/	ih	Boboi sedang gosok <i>gigi</i> <i>Boboi is brushing his teeth</i>	Gigi (<i>teeth</i>)	Gigi	g-ih-g-ih
/o/	ow	Ibu siram <i>pokok</i> bunga <i>Mother is watering the flowers</i>	Pokok (<i>tree</i>)	Pokok	p-ow-k-ow-k
/u/	uw	<i>Itu</i> gajah <i>That is an elephant</i>	Itu (<i>that</i>)	Itu	ih-t-uw
/ə/	er	<i>Kuda</i> lari laju <i>The horse ran fast</i>	Kuda (<i>horse</i>)	Kudə	k-uw-d-er

Vowels were extracted from the final syllable of the respective words, which are bola (ball), leher (neck), gigi(teeth), pokok (tree), itu (that), and kuda (horse). The selected words are disyllabic, which means that the final syllable is pronounced with prolonged vowel duration. Each vowel was segmented into 150 milliseconds. The acoustic analysis was performed using the Windows-based version of Praat software (Boersma and Weenik, University of Amsterdam, The Netherlands).

4.9.2 Word Error Rate from ASR

In deriving at the Word Error Rate (WER) for ASR system, we used the same procedures for speech training and evaluation as presented in Section 4.7.2. MFCC is used in the ASR system and the procedures in speech feature extraction are also presented in the same section. The WER is derived from 30 SIG and 30 CG speakers.

4.9.3 Statistical Analysis

The statistical analysis was performed using the Windows-based IBM SPSS Statistics 21. The analysis was conducted to determine the significant group mean differences of F1, F2, F0, intensity, jitter and shimmer between the CG and SIG using the ANOVA. In carrying out this analysis, the subject group independent variables were classified for comparison between the CG and SIG. The dependent variables are the speech features. We want to study the effect of one or more group means of speech features on the number of groups; CG and SIG. Specifically; it tests the null hypothesis (H0):

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

where μ = group mean and k = number of groups. We begin with the assumption that the H0 is TRUE, where there were significant differences in speech features between CG and SIG. Otherwise, we accept the alternative hypothesis (Ha) where at least two group means are significantly different from each other.

The statistical analysis is also conducted for the ASR performance using WER. To understand the effect of MFCC speech features of SIG speech and CG speech in relation with the ASR performance, we used the Pearson correlation to determine the correlation between WER in SIG and CG speech.

4.9.4 Results of Acoustic Analysis

This section presents the results derived from the acoustic analysis for impaired speeches and unimpaired speeches.

Formant Frequencies (Hz)

The mean values of F1 and F2 for all members of the CG and SIG are shown in Table 4.9. The overall means and standard deviations (s.d.) of F1 for CG, and SIG are 599.15 ± 80.59 , and 696.69 ± 128.28 , respectively; and the values for F2 are 1730.55 ± 274.70 , and 1644.19 ± 269.44 , respectively. Differences between F1 and F2 were found to be insignificant at $p < 0.05$, ($F = 2.088$, $p = 0.179$; $F = 0.131$, $p = 0.725$).

Table 4.9: The mean and s.d. of F1 and F2 for CG and SIG

Group	Features	Vowels						Overall mean
		/a/	/e/	/i/	/o/	/u/	/ə/	
CG	F1 (Hz)	825.80 ± 92.51	595.50 ± 75.20	462.60 ± 99.52	647.22 ± 72.29	487.22 ± 64.33	576.56 ± 79.72	599.15 ± 80.59
	F2 (Hz)	1667.43 ± 235.73	2324.59 ± 365.61	2229.95 ± 498.76	1161.32 ± 184.96	1187.86 ± 170.96	1812.15 ± 192.16	1730.55 ± 274.70
SIG	F1 (Hz)	888.94 ± 135.87	658.36 ± 115.74	622.03 ± 138.03	671.93 ± 127.08	618.39 ± 112.20	720.47 ± 140.76	696.69 ± 128.28
	F2 (Hz)	1594.53 ± 231.67	1921.50 ± 372.73	2085.20 ± 378.06	1286.20 ± 223.80	1381.29 ± 205.10	1596.41 ± 205.27	1644.19 ± 269.44

Fundamental frequency (pitch) and Intensity (dB)

Table 4.10 shows the mean and the s.d. of F0 and intensity for each group. The overall mean and s.d. of F0 and intensity for the CG and SIG are 256.04 ± 41.59 , 223.10 ± 66.03 and intensity for the CG and SIG are 60.58 ± 6.34 , 57.44 ± 8.04 , respectively.

Table 4.10: The mean and s.d. of F0 and intensity

Group	Features	Vowels						Overall mean
		/a/	/e/	/i/	/o/	/u/	/ə/	

CG	F0	257.56±	257.09±	237.85±	247.20±	275.60±	260.96±	256.04±
		37.84	46.56	51.01	44.95	28.52	40.63	41.59
	Intensity	66.26±	61.54±	50.78±	6257±	60.02±	62.29±	60.58±
		5.49	6.28	5.31	5.88	9.17	5.93	6.34
SIG	F0	208.51±	245.82±	215.94±	211.92±	229.79±	226.60±	223.10±
		66.32	56.76	72.81	59.09	64.99	76.19	66.03
	Intensity	59.60±	60.21±	51.79±	58.42±	56.37±	58.23±	57.44±
		7.37	7.70	8.54	7.13	7.20	10.32	8.04

There were significant differences between CG and SDG in F0 at $p < 0.05$ ($F = 18.279$, $p = 0.002$), while intensity were found insignificant differences between CG and SIG at $p < 0.05$ ($F = 1.613$, $p = 0.233$).

Jitter and shimmer

Table 4.11 summarises the means of jitter and shimmer for all groups. For the CG, the mean and s.d. values of jitter and shimmer are 0.63 ± 0.34 , 3.78 ± 1.64 and 14.47 ± 5.22 , respectively. For the SIG, the mean and s.d. values of jitter and shimmer are 1.78 ± 1.43 , 8.78 ± 4.53 and 11.90 ± 4.98 , respectively.

Table 4.11: The mean and s.d. values of jitter and shimmer for the CG and SIG

Group	Features	Vowel						Overall mean
		/a/	/e/	/i/	/o/	/u/	/ə/	
CG	Jitt (%)	0.48±0.32	0.51±0.40	0.54±0.27	0.65±0.31	0.87±0.40	0.72±0.4	0.63±0.34
	Shim (%)	3.04±1.34	3.67±1.76	4.25±1.90	4.02±1.87	3.88±1.50	3.79±1.5	3.78±1.64
SIG	Jitt (%)	1.55±1.41	1.37±0.91	2.18±1.51	2.00±1.71	1.72±1.41	1.8±1.6	1.78±1.43
	Shim (%)	8.48±4.56	7.25±3.58	9.34±4.33	9.73±4.30	8.11±4.71	9.8±5.7	8.78±4.53

Jitter and shimmer values for the CG is much lower compared with the SIG. It shows that fewer perturbation values are found in the speech of normal children compared with the speech of impaired-speech children. There are significant differences between at $p < 0.05$ for the ratings of the CG and SIG in jitter ($F = 71.894$ $p = 0.000$) and shimmer, ($F = 125.830$, $p = 0.000$).

Differences between CG and SIG

Table 4.12 concludes the differences between CG and SIG for each speech features.

Table 4.12: Differences between CG and SIG for each features

Features	CG	SIG	Mean difference	P-value
F1	599.15	696.69	97.54	0.179
F2	1730.55	1644.19	86.36	0.725
F0	256.04	223.10	32.94	0.002*
Intensity	60.58	57.44	3.14	0.233
Jitter	0.63	1.78	1.15	0.000*
Shimmer	3.78	8.78	5.00	0.000*

* $p < 0.05$

Figure 4.15 shows the values changes in SIG speech when compared with CG. There are increment in F1, Jitter and Shimmer. There is a high increment of Jitter and Shimmer in SG, which are 1.15% (182%) and 5.00% (132%), respectively. F1 for SIG increases 97.54Hz (16%). Meanwhile, F2, F0 and Intensity reduce in SIG speech, 86.36Hz (5% reduction), 32.94Hz (13% reduction) and 3.14 (5% reduction) respectively. Overall, we can conclude that there are differences in speech feature values between the SIG and CG.

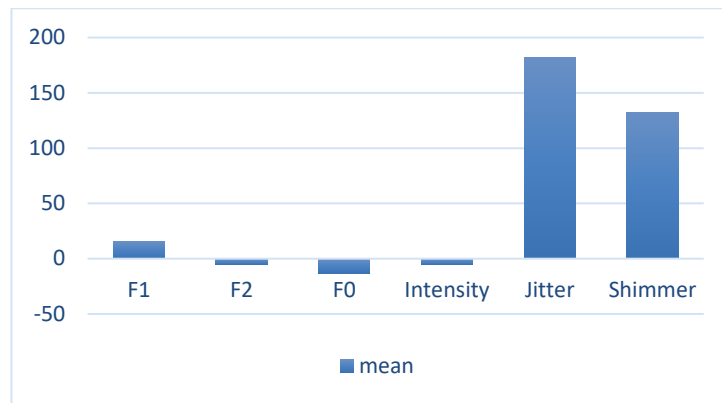


Figure 4.15: Mean differences

4.9.5 Results of ASR performance

Figure 4.16 shows the difference of WER where SIG was 39.23% higher than CG. There is a strong negative correlation, which the WER increase with the decrease of PCC value or the degree of severity impairment. The correlation between severity of impairments with WER is significant at $p < 0.05$, ($r = -0.95$, $p = 0.00$). It is shown that WER for SIG speakers increase with the decrease of intelligibility in speech.

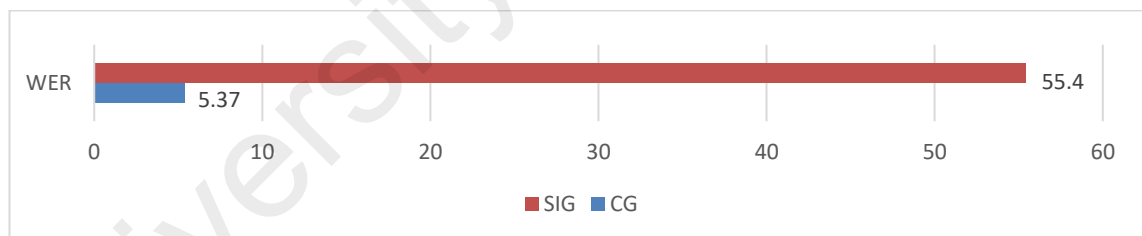


Figure 4.16: Comparison of WER between SIG and CG

4.9.6 Summary of findings

The previous section presented the results obtained and each of the findings is discussed as below.

Acoustic analysis

We have identified the speech features that contribute to the intelligibility deficits in impaired speech among children. F0, jitter and shimmer were found to show significant differences for impaired speech.

- F0

In this study, the statistical analysis shows that there is significant difference in F0 between the CG and SIG. The results of the F0 reduction in SIG are in agreement with (Jeng et al., 2006), who claimed that F0 tend to be lower for impaired speech. However, some studies have reported that there is no significant difference in F0 decrement in impaired speech (Saz et al., 2009a). This is because speakers of impaired speech can still control some prosodic features in their speech, even though they lose intelligibility in vowel production (Saz et al., 2009a; Patel, 2002).

- Jitter and shimmer

SIG speakers have higher jitter and shimmer compared with the CG. The statistical analysis shows that there is a significant difference in jitter and shimmer between the CG and SIG, which is in agreement with (Hartl et al., 2003). However, our findings contradict those of other studies (Wertzner et al., 2005) which claimed that there are no differences in jitter and shimmer between impaired and normal children.

Overall, the acoustic analysis revealed that F0, jitter and shimmer are significant features in contributing to low intelligibility of Malay children with impaired speech.

Table 4.13: Comparison of findings in acoustic analysis of impaired speech

Author	Features studied						
	Language	Formant	F0	Intensity	Jitter	Shimmer	Duration
Author, 2016	Malay	NS	S	NS	S	S	-

White, 2012	English	-	NS	-	NS	NS	-
Saz et al., 2009a	Spanish	S	NS	NS	-	-	NS
Jeng et al., 2006	Mandarin	-	S	-	-	-	-
Wertzner et al., 2005	Portuguese	-	S for vowel /e/	-	NS	NS	-
Hartl et al., 2003	French	-	-	-	S	S	-

S=Significant, NS=Not Significant

Table 4.13 shows comparison of these research findings with the literature. The results of acoustic analysis for Malay impaired speech data were found to be similar with Mandarin language for F0 and Spanish for intensity, respectively. On the other hand, jitter and shimmer are similar to French. These similarities happen due to the characteristics of pronunciations for the particular language which affect the particular speech features.

4.10 Summary

This chapter presents the process of developing the speech corpus, measuring the speech intelligibility and analyzing the recorded speeches. The important outcomes are as follows;

- A corpus of Malay speaking children with speech impairments is developed.
- The speech intelligibility scores are measured by using subjective measurement by SLPs and automatic measurement using ASR system. Even though ASR system was suggested as one way of measuring intelligibility, it was found that the intelligibility scores are not as reliable as compared to the scores of SLPs. This indicates that the importance of incorporating the speech knowledge such as speech features to detect

the abnormal variation in impaired speeches in order to give reliable scores and classify intelligibility. Detection based ASR is used in this research for classifying the speech intelligibility rather than using ASR only.

- For the acoustic and statistical analysis conducted, it was found that there are significant differences between non-impaired and impaired speech.

University of Malaya

CHAPTER 5 AUTOMATIC SPEECH INTELLIGIBILITY DETECTION FOR MALAY SPEECH IMPAIRED CHILDREN: A BASELINE RESULT

This chapter presents the development of the automatic speech intelligibility detection for Malay impaired speakers. This section explains each steps and methods that are carried out in developing the automatic speech intelligibility detection using the selected baseline classification methods which are Support Vector Machine (SVM), Random Forest (RF), Linear Discriminant Analysis (LDA) and k-Nearest Neighbour (KNN).

5.1 Automatic Speech Intelligibility Detection for Malay Impaired Speakers

This section presents the automatic intelligibility detection for impaired speech system using baseline classifiers as identified in **Chapter 3**. There is a consensus in literature in **Chapter 2** and experimental analysis in **Chapter 4** to understand the speech features that are salient in reducing the speech intelligibility in impaired speech. F0, Energy, ZCR, MFCC, jitter and shimmer are identified as significant speech features that correlates with the intelligibility deficits. In this section, the speech features are evaluated for the automatic detection of speech intelligibility. The speech intelligibility detection is treated as a binary classification problem, classifying words as either intelligible or not intelligible. Figure 5.1 shows the general framework of the automatic speech intelligibility detection. Basically, the speech intelligibility detection consists of the speech data, speech feature extraction, classification methods, training and evaluation phase. Speech signals $s(t)$ are inputted to the speech feature extraction to extracts the meaningful and significant speech features $On(t)$ in the classification task. Then, the training and evaluation phase are performed that make use of the knowledge

scores $P(A_n|O_n(t))$ produces by the classification methods in order to detect the speech intelligibility of impaired speakers. Further discussions on the implementation details are discussed in the next sections.

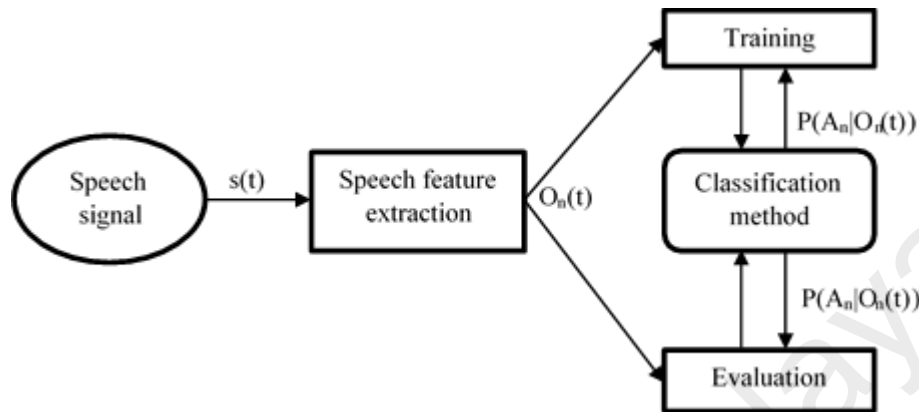


Figure 5.1: General framework of the speech intelligibility detection

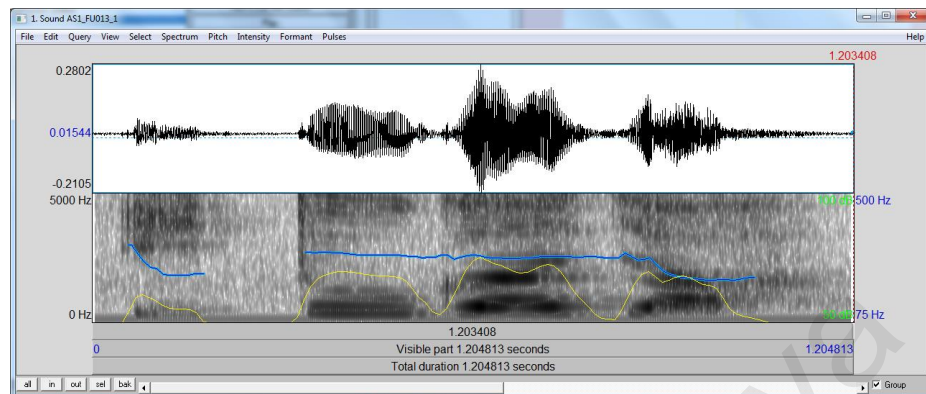
5.1.1 Data Preparation

The initial steps in automatic speech intelligibility detection is to prepare the speech data which will later be used for training and evaluation. The experiments presented are trained and evaluated on the speech corpus of Malay impaired speakers and the control group. This development of the speech corpus is described thoroughly in **Chapter 4**.

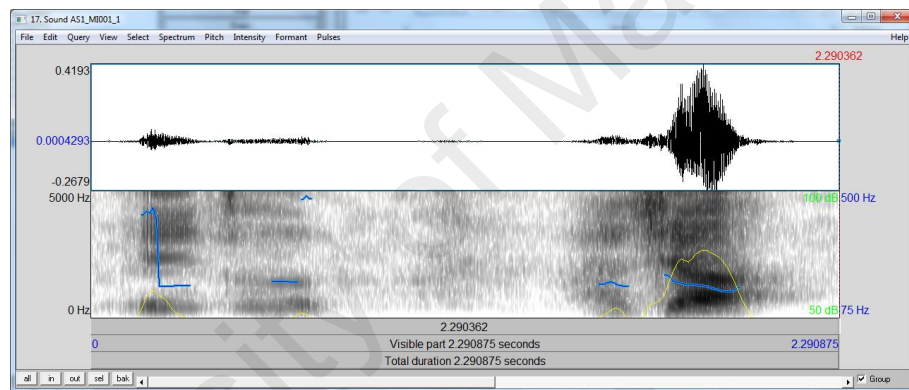
5.1.2 Speech Feature Extraction

Feature extraction for the detection uses the Opensmile toolkit. The WAVE signal of speech utterances are segmented into 25 milliseconds with an interval of 10 milliseconds. Feature extraction is performed to analyze the non-stationary behaviour of the audio samples of speech signal to detect the speech intelligibility. The speech is parameterized with 13 MFCC (0th to 12th coefficients), normalized log energy, ZCR, F0, jitter local and shimmer local yielding a total of 18 dimensional feature vector. The feature vectors are then used for further process for the classifier. Figure 5.2 shows the

example of F0 contour, intensity and spectral of (a) intelligible and (b) not intelligible for the same sentence; “*itu gajah*”.



(a) intelligible speech



(a) not intelligible speech

Figure 5.2: Example of pitch contour of (a) intelligible speech and (b) not intelligible speech for sentence “*itu gajah*”

In Opensmile, the classification according to speech intelligibility is assigned using the variable `class.type` in which includes both intelligible and notintelligible classes. In this step, the files belong to intelligibility class has been assigned to *intelligible* and *notintelligible* as follows:

```
class[0].type = { intelligible, notintelligible }; default class
target[0].all = intelligible
```

```
class[0].type = { intelligible, notintelligible }; default class
target[0].all = notintelligible
```

The speech feature extraction files are generated separately in .arff files for both training and evaluation.

5.1.3 Baseline Classification Methods

This section describes the set-up for baseline classification methods to detect the speech intelligibility of impaired speech. LibSVM Matlab toolbox is used for SVM model training and evaluation. The randomForest package is used for training and evaluating the RF which is a MATLAB standalone application. For LDA and KNN, the default parameter provided by MATLAB used without any modification.

5.2 Evaluation of Baseline Classification Methods

The evaluation is carried out on the four selected classifiers, SVM, RF, LDA and KNN, which implemented in MATLAB 2013b. The 10 fold cross validation is used where the speech files were randomly partitioned into 10 equal size subsamples, where nine partitions are set for the training to train the model and the remaining one partition is the test set for evaluating the model. In each run, one of the partition is used as a test data and the remaining partitions are used as train data. This procedure is repeated 10 times until all 10 subsamples are used as test data. The performance of classifiers are the average of training and testing data.

The evaluation of the baseline system involves speeches from 30 CG and 30 SIG speakers. A total of 2,950 utterances from 1,528 unimpaired utterances and 1,422 impaired utterances were used for the evaluation purposes. These utterances were extracted using Opensmile to get the significant speech features; energy, f0, zcr, mfcc

0th coefficient to 12th coefficient, jitter and shimmer. The extracted speech features used are shown in (*Appendix C: Speech Data for Classification*)

The result presents for baseline classifiers are measured using the Classification Accuracy, Precision and Recall. The confusion matrix of classification error rate for Type I and Type II are presented as well. There are two types of evaluation performed. First, the speech data are evaluated in terms of classification rate, classification accuracy, precision and recall for the overall data. Second, the classification accuracy is derived for individual speech feature.

5.3 Result

Table 5.1 presents the confusion matrix of misclassification for SVM, RF, LDA and KNN. RF produces the highest Type 1 (FP) with 10 times, follows by SVM (6), KNN (3) and LDA (2). Meanwhile, for Type II error (FN), SVM and LDA produces the highest which are 18 frequencies, follows by KNN (15), and RF (7).

Table 5.1: Confusion matrix of the baseline classification methods

SVM		Prediction		Total
		Notint	Int	
Actual	Notint	TP(147)	FP(6)	153
	Int	FN(18)	TN(124)	142
	Total	165	130	295
LDA		Prediction		Total
		Notint	Int	
Actual	Notint	TP(151)	FP(2)	153
	Int	FN(18)	TN(124)	142
	Total	169	126	295
RF		Prediction		Total
		Notint	Int	
Actual	Notint	TP(143)	FP(10)	153
	Int	FN(7)	TN(135)	142
	Total	150	145	295
KNN		Prediction		Total
		Notint	Int	
Actual	Notint	TP(150)	FP(3)	153
	Int	FN(15)	TN(127)	142
	Total	165	130	295

Table 5.2 shows the classification accuracy results of the baseline classifiers for all the 10 folds. The accuracy values are rounded to two decimal places. Later, the average of training and testing accuracy are calculated as the performance of the classifier.

Table 5.2: Classification accuracy of the baseline classification methods for each fold

Fold	SVM		RF		LDA		KNN	
	training	testing	training	testing	training	testing	training	testing
Fold 1	96.99	96.28	99.95	96.31	96.91	95.93	100	98.98
Fold 2	96.88	94.93	98.06	94.97	96.65	93.83	100	98.64
Fold 3	97.1	94.59	99.89	94.64	97.25	93.92	100	98.98
Fold 4	96.91	97.29	99.85	95.37	96.76	96.95	100	97.62
Fold 5	96.59	98.64	99.76	95.56	96.73	95.96	100	97.63
Fold 6	97.19	96.63	99.64	96.39	97.18	93.77	100	97.29
Fold 7	96.79	96.32	99.1	95.37	96.84	95.95	100	97.29
Fold 8	95.69	96.98	99.42	94.17	96.8	96.81	100	99.66
Fold 9	96.62	95.75	99.97	96.17	96.64	96.95	100	97.97
Fold 10	96.32	96.95	98.71	94.69	96.95	97.43	100	93.90
Average	96.71	96.44	99.44	95.36	96.87	95.75	100	97.80

Table 5.3 presents the classification accuracy, precision and recall of the selected baseline classifiers in terms of the training and evaluation. The classification accuracy training set for SVM is 96.71%, RF is 99.40% and LDA is 96.87%. KNN produces 100% classification accuracy. In evaluation, KNN produce the highest accuracy with 97.80%, follows by SVM with 96.44%. LDA produces 95.75% and RF with slightly which is 93.22%. Meanwhile, LDA produces the highest precision with 98.62%,

follows by KNN with slightly lower, 98.53%, SVM (96.34%) and RF (93.23%). For recall, RF produces highest percentage with 95.67%, follows by KNN, SVM and LDA with 90.90%, 89.65% and 89.32%, respectively.

Table 5.3: The overall classification accuracy, precision and recall of baseline classifiers

Classification method	Accuracy		Precision		Recall	
	Training	Evaluation	Training	Evaluation	Training	Evaluation
SVM	96.71	96.44	98.08	96.34	96.72	89.65
RF	99.40	95.36	99.08	93.23	99.51	95.67
LDA	96.87	95.75	99.15	98.62	96.01	89.32
KNN	100.00	97.80	100.00	98.53	100.00	90.90

Table 5.4 shows the accuracy for all the baseline classification methods for each individual features. Mean values are calculated for each aspect of speech features such as prosody, pronunciation and the voice quality aspect.

Table 5.4: The classification accuracy based on the individual speech features

Speech features	SVM		RF		LDA		KNN	
	Training	Evaluation	Training	Evaluation	Training	Evaluation	Training	Evaluation
Prosody								
F0	56.67	56.73	76.92	70.08	61.59	61.16	73.22	67.23
Energy	51.77	51.71	75.17	74.69	51.59	51.81	56.30	56.35
ZCR	80.58	80.58	86.03	78.89	80.67	80.58	86.00	78.55
Mean	63.01	63.01	79.37	74.55	64.62	64.52	71.84	67.38
Pronunciation								

MFCC 0	80.90	80.89	85.17	77.33	80.94	80.92	82.49	73.77
MFCC 1	61.62	61.60	71.75	57.30	61.62	61.74	60.50	50.53
MFCC 2	83.25	83.19	88.58	79.67	83.20	83.19	88.50	78.42
MFCC 3	86.05	86.04	87.14	84.34	86.20	86.17	76.93	74.72
MFCC 4	72.95	73.16	85.08	71.16	72.89	72.86	85.85	71.06
MFCC 5	79.99	79.97	85.58	77.40	80.08	80.07	85.05	75.84
MFCC 6	59.49	59.57	78.78	56.39	59.07	59.03	78.82	55.30
MFCC 7	75.02	74.96	84.83	68.62	75.02	74.96	85.61	66.96
MFCC 8	69.60	69.57	81.74	63.94	69.31	69.37	81.16	61.71
MFCC 9	81.20	81.23	87.68	77.60	81.18	81.23	88.18	75.23
MFCC 10	72.16	72.14	83.12	66.15	71.87	71.87	84.12	64.59
MFCC 11	73.35	73.43	84.30	69.61	73.34	73.37	84.22	66.66
MFCC 12	73.15	73.20	74.18	71.60	73.22	73.13	59.04	58.39
Mean	74.52	74.53	82.92	70.85	74.46	74.45	80.04	67.17
Voice quality								
Jitter	73.26	73.50	86.45	75.09	76.33	76.38	86.44	74.21
Shimmer	78.90	78.99	87.75	79.02	82.13	82.04	87.07	77.57
Mean	76.08	76.25	87.10	77.06	79.23	79.21	86.76	75.89

For prosody, the mean value of classification accuracy for RF is the highest at 74.55%, KNN at 67.38%, followed by LDA and SVM, at 64.52% and 63.01%, respectively. In term of pronunciation, SVM gained the classification accuracy highest mean values at 74.53%, followed by LDA at 74.45%. On the other hand RF and KNN at 70.85% and 67.17%, respectively. For voice quality, LDA has the highest mean values of classification accuracy at 79.21%, followed by RF, SVM and KNN at 77.06%, 76.25% and 75.89%, respectively.

5.4 Discussion on Findings

Based on the classification results of baseline classifiers, the discussion of findings are as follows;

5.4.1 Accuracy, Precision and Recall of Baseline Classification Methods

Based on the classification results as presented in Table 5.2, different classifiers perform differently in term of classification accuracy, precision and recall. For classification accuracy, KNN has the highest score among the four classifiers (97.80%).

LDA has the highest score for precision (98.62%), while for recall, RF has the highest score (95.67%). It was also found that the classifiers performance during evaluation degrades as presented in Table 5.5. For Classification accuracy and Precision, RF shows the highest degradation of 4.04% point and 5.85% point, respectively. SVM has the highest degradation for recall of 7.07% point.

Table 5.5: The degradation values of training and evaluation for each baseline methods

Classification method	Accuracy		Degradation	Precision		Degradation	Recall		Degradation
	Training	Evaluation		Training	Evaluation		Training	Evaluation	
SVM	96.71	96.44	0.27	98.08	96.34	1.74	96.72	89.65	7.07
RF	99.40	95.36	4.04	99.08	93.23	5.85	99.51	95.67	3.84
LDA	96.87	95.75	1.12	99.15	98.62	0.53	96.01	89.32	6.69
KNN	100.00	97.80	2.2	100.00	98.53	1.47	100.00	90.90	9.1

Figure 5.4 compares the accuracy, precision and recall of the baseline classification methods. For accuracy, all classifiers produce almost similar percentage where the lowest is RF with 95.75% and the highest is KNN with 97.80%. It seems that precision

produces the promising measures for classifying impaired speech as it has higher values for all classification methods except RF. However, for the recall, all classification methods produce lower values except RF.



Figure 5.3: Comparison of the accuracy, precision and recall for baseline classification methods

Table 5.6 shows the mean values for accuracy, precision and recall for all baseline classification methods. Precision has the highest mean value of 96.68%, follows with accuracy at 96.34% and recall at 91.39%. Accuracy is the most common performance measure in classification where it is simply the ratio of correctly predicted observations.

Table 5.6: Mean values for the accuracy, precision and recall

Classification method	Accuracy	Precision	Recall
SVM	96.44	96.34	89.65
RF	95.36	93.23	95.67
LDA	95.75	98.62	89.32
KNN	97.80	98.53	90.90
Mean value	96.34	96.68	91.39

Table 5.7 shows the mean values of accuracy, precision and recall for each baseline classification methods. In terms of the average accuracy, precision and recall evaluation, KNN has the highest mean value at 95.74%. On the other hand, RF, LDA and SVM produce almost similar mean values at 94.75%, 94.56% and 94.14%.

Table 5.7: The mean values of accuracy, precision and recall for each baseline methods

Classification method	Accuracy	Precision	Recall	Mean value
SVM	96.44	96.34	89.65	94.14
RF	95.36	93.23	95.67	94.75
LDA	95.75	98.62	89.32	94.56
KNN	97.80	98.53	90.90	95.74

Among the classification methods, RF is observed to produce different or not synchronize measures especially for precision and recall. This happens due to its characteristics of RF model which was treated as a black box. Therefore, gaining a full understanding of decision process by examining each individual tree is not feasible. In fact, for most of the statistical or machine learning approach such as SVM, use the black box approach in classification. These methods have been proven to be effective in terms of predictive accuracy, but they provide little meaningful explanation of prediction and give little new insight about the data or the application domain (Freitas et al., 2010).

5.4.2 The relation of speech features with speech intelligibility classification

Based on Table 5.2 and 5.3, it is obvious that the combination of selected speech features have more discriminating power that produces higher classification accuracy compared to the individual speech features and the mean values for each aspect of speech.

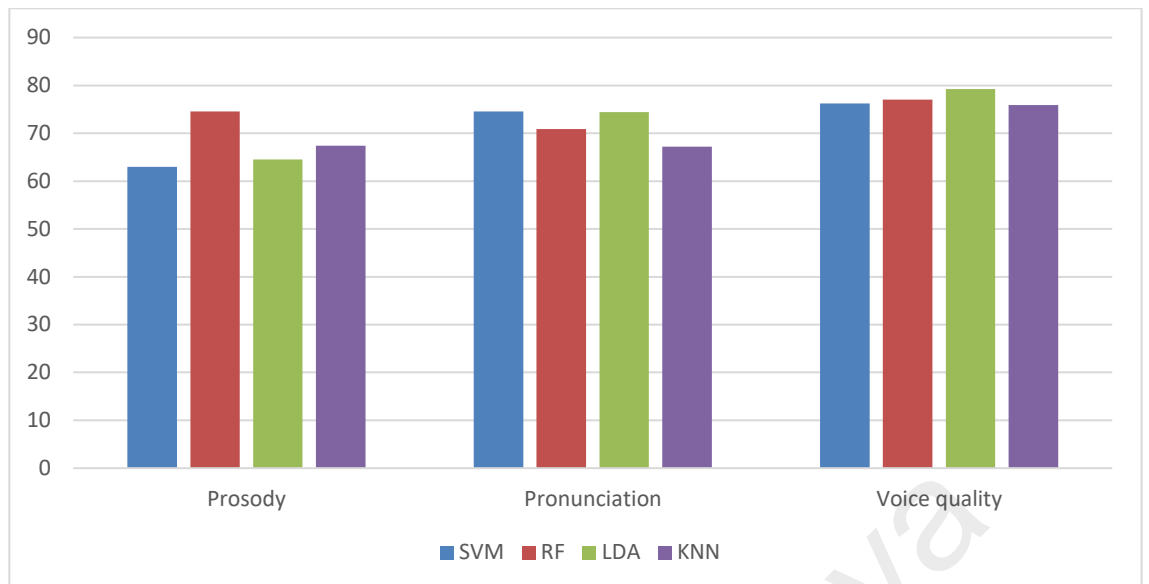


Figure 5.4: The graph comparing the accuracy based on the prosody, pronunciation and voice quality

In Figure 5.3, voice quality indicates the highest accuracy for all classification methods. This result correlates to the findings of speech analysis in **Section 4.9** which indicates that jitter and shimmer are significant speech features that contributes to the speech intelligibility deficits among impaired speakers.

5.5 Summary

This chapter presents baseline classification methods to obtain the benchmark results of the classification accuracy, precision and recall. We have also presented the classification accuracy for each individual speech features with the mean values of the three aspect of speech includes the prosody, pronunciation and voice quality. Further discussion of the performance of benchmark classification methods are presented in the findings section.

CHAPTER 6 THE PROPOSED FUZZY PETRI NETS (FPN) CLASSIFICATION METHOD FOR SPEECH INTELLIGIBILITY DETECTION

This chapter presents the proposed FPN as classification method for speech intelligibility detection. It consists of the proposed FPN framework, the development procedures and the evaluation that leads to the performance comparison with the benchmark classification methods as discussed in **Chapter 5**. Summary of the proposed classification method then is discussed.

6.1 Proposed Approach

This section presents the proposed FPN as classification method for speech intelligibility detection.

There is a consensus in **Chapter 3** to identify a suitable classification method for addressing the issues of intelligibility detection performance in terms of accuracy, precision and recall. The significant speech features that correlates with the speech intelligibility of impaired speech identified in **Chapter 3 and 4**. Later, individual speech feature is evaluated on the proposed FPN to understand the relationship between the speech features and the performance of the intelligibility detection. Selected speech features are evaluated for automatic detection of speech intelligibility, which will be later evaluated using the Malay speech database of impaired speakers and the control group. Figure 6.1 shows the framework of the proposed FPN for the automatic speech intelligibility detection.

- Speech feature extraction to extract the discriminative features

- FPN classification that consists of the Fuzzy Inference System (FIS) and FPN Modelling

The proposed approach used similar procedures of data preparation and speech feature extraction as described in **Section 5.1**. The development of FIS and FPN modeling in the FPN classification for detecting the speech intelligibility are discussed further in the next sections.

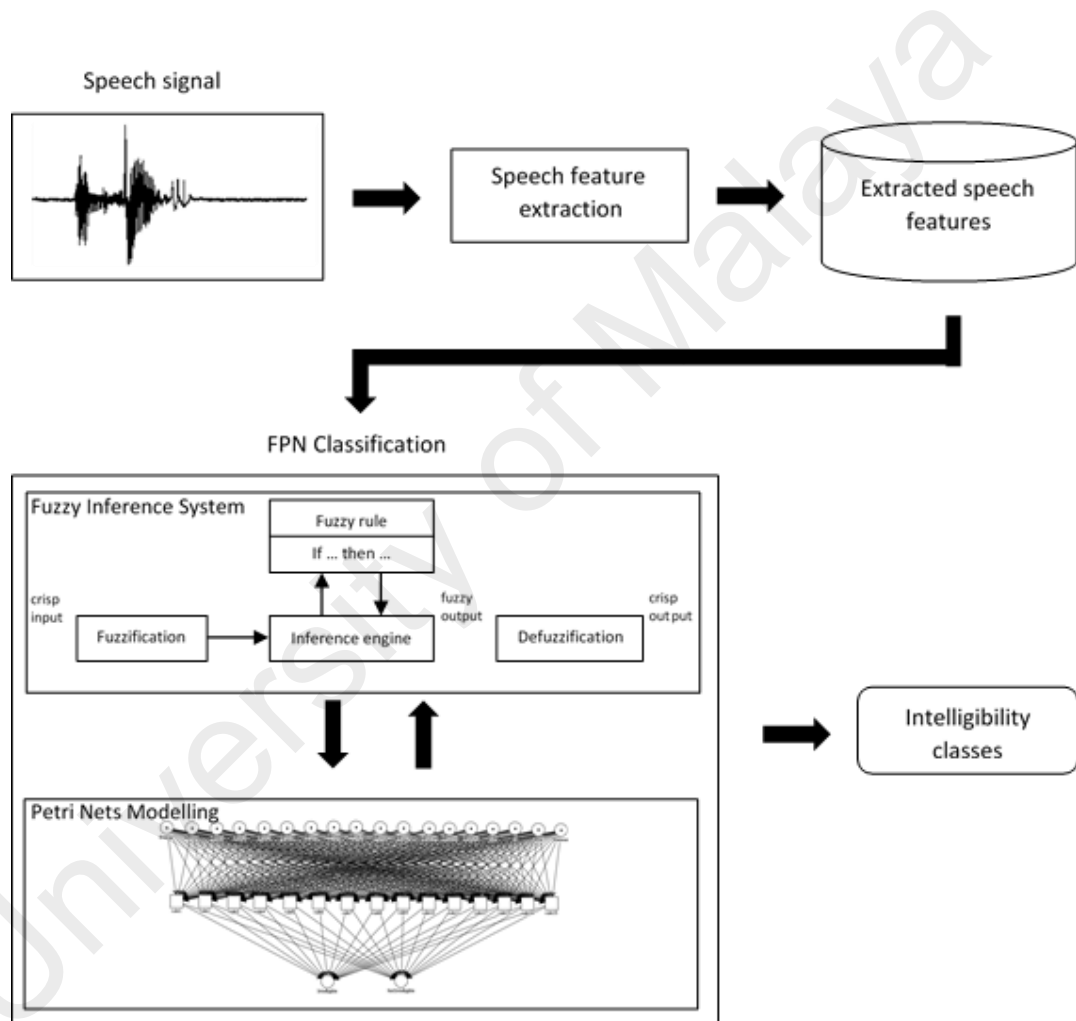


Figure 6.1: The framework of the proposed FPN

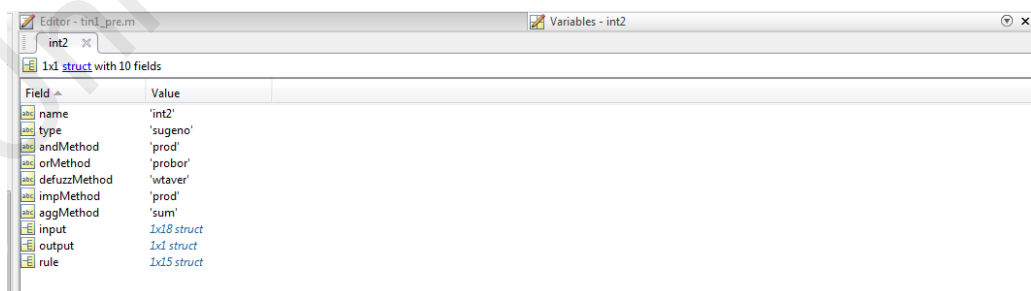
6.2 Fuzzy Inference System (FIS)

As discussed in Chapter 3, Subtractive Clustering is used to generate the FIS. A FIS structure is needed to perform the training for the FIS. In this research, `genfis2` is used for training the FIS model;

```
fismat = genfis2(input,output);
```

It generates a FIS by extracting a set of rules that models the data behavior. This function requires separate sets of input and output data as input arguments. For input data, we have speech features comprising MFCC value of 0th to 12th coefficients, energy, zcr, f0, jitter and shimmer which in total has 18 attributes for input. On the other hand, output has 1 attribute as either intelligible (1) or not intelligible (2). When there is only one output, `genfis2` can be utilized to generate an initial FIS for ANFIS training. The rule extraction method first uses the clustering function to determine the number of rules and membership functions for the antecedents and consequents. It produces the iteration count for training the FIS (*Appendix D: Iteration Count of FIS*).

The Sugeno-type FIS structure (`fismat`) input data, `input` and output data, `output` is shown in figure 6.2.



Field	Value
name	'int2'
type	'sugeno'
andMethod	'prod'
orMethod	'probor'
defuzzMethod	'wtaver'
impMethod	'prod'
aggMethod	'sum'
input	1x18 struct
output	1x1 struct
rule	1x15 struct

Figure 6.2: `fismat` structure

A total of 15 clusters were derived where the number of clusters determines the number of rules and membership functions from the generated FIS as shown in Figure 6.3.

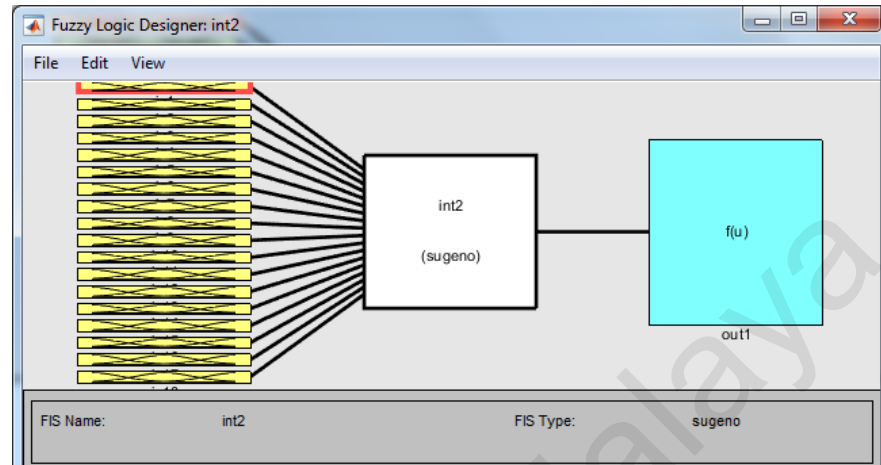


Figure 6.3: Fuzzy Inference System

The input membership function type is 'gaussmf'. Figure 6.4 shows the 15 membership functions plot in Gaussian.

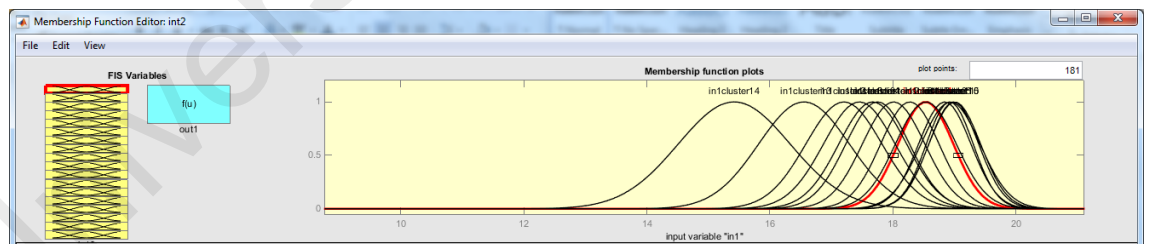


Figure 6.4: Gaussian membership function plots

6.2.1 Fuzzy rule based reasoning and FPN classification

15 rules were generated from FIS using the subtractive clustering. The following is the example of the first rule generated from the 15 rules. All 15 rules generated are shown in (*Appendix E: Fuzzy Rules and Fuzzy Rules Viewer*).

Rule 1

If (*in1 is in1cluster1*) and (*in2 is in2cluster1*) and (*in3 is in3cluster1*) and (*in4 is in4cluster1*) and (*in5 is in5cluster1*) and (*in6 is in6cluster1*) and (*in7 is in7cluster1*) and (*in8 is in8cluster1*) and (*in9 is in9cluster1*) and (*in10 is in10cluster1*) and (*in11 is in11cluster1*) and (*in12 is in12cluster1*) and (*in13 is in13cluster1*) and (*in14 is in14cluster1*) and (*in15 is in15cluster1*) and (*in16 is in16cluster1*) and (*in17 is in17cluster1*) and (*in18 is in18cluster1*) then (*out1 is out1cluster1*) (1)

where *in1* represent energy, *in2* to *in14* are MFCC0 to MFCC12, *in15* is ZCR, *in16* is F0, *in17* is jitter and *in18* is shimmer. Meanwhile, *cluster1* is represents speech features in the cluster of intelligibility.

The FPN reasoning is described in Figure 6.5. The 15 rules are represented with 15 transitions (*tin1, tin2, tin3 tin15*), with 15 input arcs and 1 output arc. There are 18 speech features as input variables. As such, FPN contains two fuzzification transitions (*fuz-in1, fuz-in2, fuz-in3, fuz-in15*) which read the input and maps them to corresponding places (*in1, in2, in3,, in15*) in the knowledge base and decision making logic. For the 2 output variable (*intelligible* and *notintelligible*), a defuzzification transition (Def-intelligible, Def-notintelligible) is used to create the output signal to the system being controlled. The labels on the arcs are labels of linguistic values and crisp values. In our case, the defuzzification module maps the speech features to intelligibility classes for the corresponding variables.

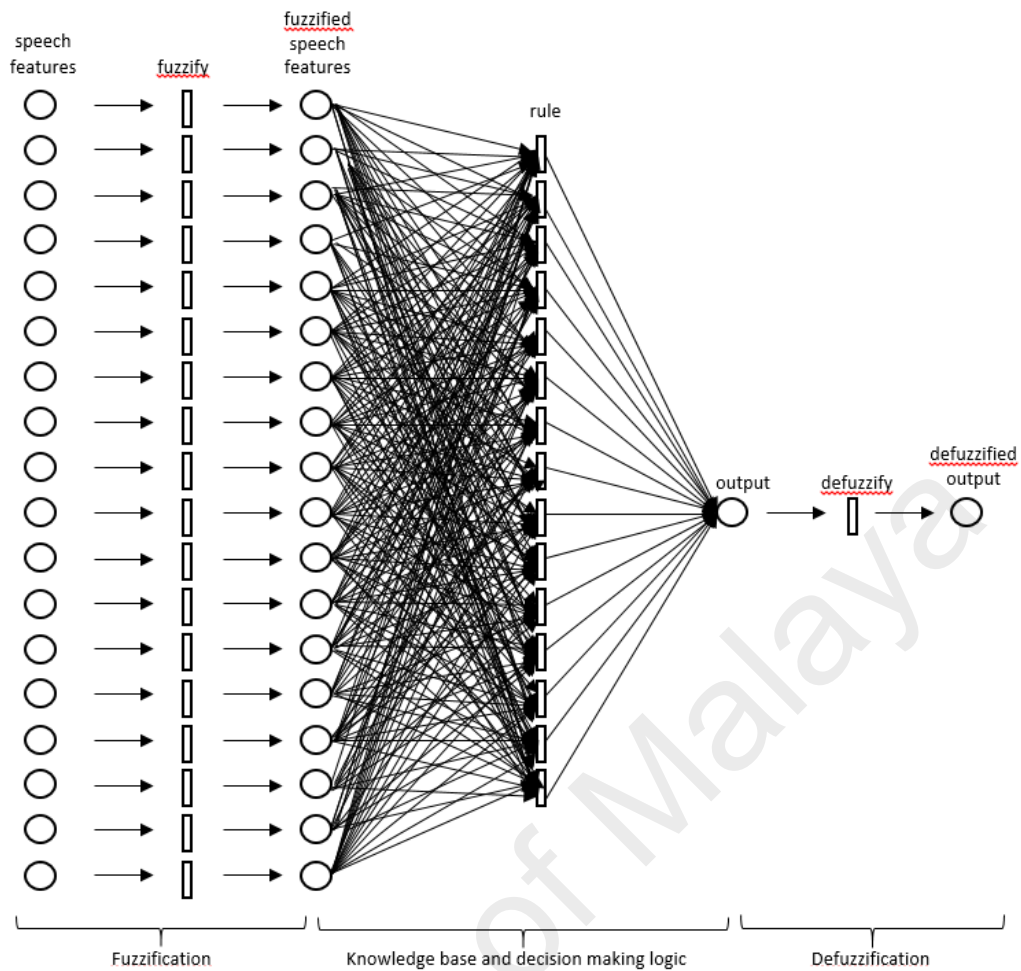


Figure 6.5: The FPN reasoning

6.3 The FPN Implementation

In previous section, the FIS for the intelligibility detection has been developed. Next, the implementation of FPN classification method is discussed in this section. Figure 6.6 shows the process flow of the PN Editor, PNML-2-GPenSIM converter and GPenSIM. These selected tools are discussed in section 3.8.2. First, PN Editor tool is used for Petri Nets modelling of intelligibility detection. This Petri Nets model is exported to Petri Net Markup Language (PNML) file, which later is converted to readable or compatible format to GPenSIM files using PNML-2-GPenSIM converter. These files are the Main Simulation File (MSF), Petri Definition file (PDF) Transition Definition File (TDF).

Finally, GPenSim is used to communicate with the FIS engine for performing the classification tasks for intelligibility detection.

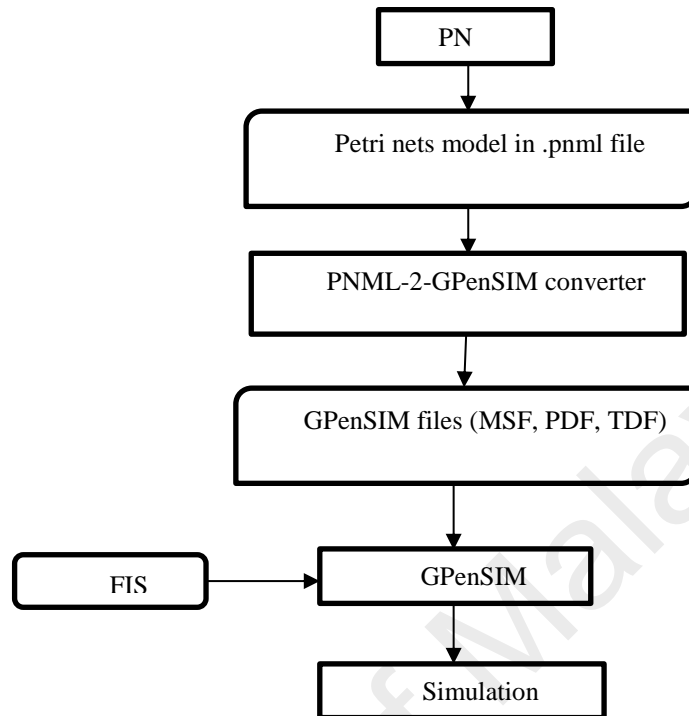


Figure 6.6: The process flow of implementation

6.3.1 The Petri Nets Modeling

The Petri Nets modeling is performed using PN Editor tool. Figure 6.7 shows the Petri Nets model that was designed of 18 tokens, representing the 18 speech features, 15 transition which represents the 15 fuzzy rules and 2 output classes of speech intelligibility which are *intelligible* or *not intelligible*.

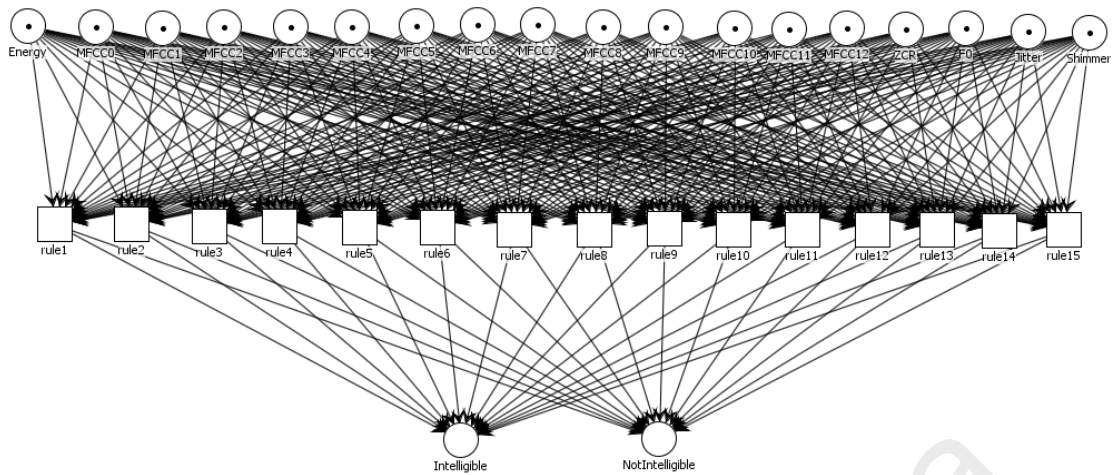


Figure 6.7: The PN Modelling using PN Editor

6.3.2 PNML to GPenSIM files

The conversion of Petri Nets model pnml file to gpenSIM files uses the PNML-2-GPenSIM converter. The pnml document is shown in *Appendix F: The pnml document*. The converter is a Matlab function that reads a PNML files of Petri Nets model, extracts the Petri Nets structure and creates PDF, MSF and TDF files representing the model. During this process, the graphical details coded in the PNML file are discarded. The function used is as follows;

```
pnml2gpenSIM(PNMLFilename);
```

The PNML-2-GPenSIM converter is built on MATLAB platform. MATLAB offers a set of functions for reading and interpreting XML files, starting with the function ‘xmlread’ that reads an XML document and returns Document Object Model (DOM) node. From the DOM node, the elements of the node (such as ‘place’, ‘transition’ and ‘arc’) can be visited recursively, extracting the names of the elements, the initial marking (in case of place element), the source and the target (in case of an arc element).

The converter generated 1 MSF, 1 PDF and 18 TDF files as shown in *Appendix G: The MSF, PDF and TDF files*.

6.3.3 GPenSIM Implementation

In GPenSIM, three files are used, the MSF, PDF and TDF. In this research, MSF, PDF and TDF are generated from the PNML-2-GPenSIM converter. Figure 6.8 shows the integration of the three types of files. The MSF file are the main .m file that is responsible to run the simulations, while PDF contains the implementation detail of a Petri Nets and TDF contains the implementation details of transitions.

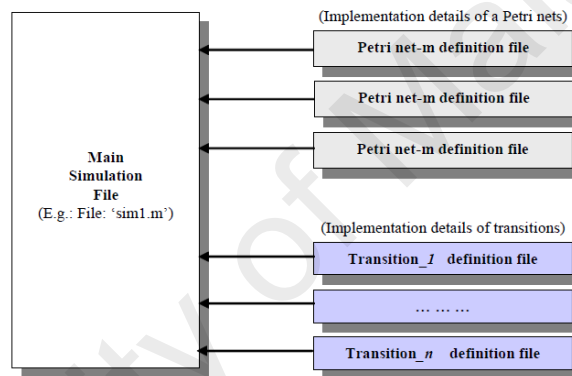


Figure 6.8: Integration of the MSF, PDF and TDF files (Davidrajuh, 2012)

The MSF consists of the following information;

- The initial markings on the places
- The initial dynamics includes the initial markings and firing times

The PDF consists of the following information;

- The places
- The transitions
- The arcs for connecting the places and transition

The TDF consists of the following information;

- TDF_PREs and TDF_POSs for the transitions, which are the user-defined conditions attached to the transitions.

6.4 Evaluation of the Proposed FPN

The speech features used for evaluating the proposed FPN are similar to the evaluation of the baseline classification methods in Section 5.2.

6.5 Results

To figure out the misclassifications produced by FPN classifier, the confusion matrix of FPN is compared to the confusion matrix of the baseline classifiers, SVM, RF, LDA and KNN.

Table 6.1: Confusion matrix of the proposed FPN and the baseline classification methods

FPN		Prediction		Total
		Notint	Int	
Actual	Notint	TP(153)	FP(0)	153
	Int	FN(9)	TN(133)	142
	Total	162	133	295
RF		Prediction		Total
		Notint	Int	
Actual	Notint	TP(143)	FP(10)	153
	Int	FN(7)	TN(135)	142
	Total	150	145	295
KNN		Prediction		Total
		Notint	Int	
Actual	Notint	TP(150)	FP(3)	153
	Int	FN(15)	TN(127)	142
	Total	165	130	295
SVM		Prediction		Total
		Notint	Int	
Actual	Notint	TP(147)	FP(6)	153
	Int	FN(18)	TN(124)	142
	Total	165	130	295
LDA		Prediction		Total
		Notint	Int	
Actual	Notint	TP(151)	FP(2)	153
	Int	FN(18)	TN(124)	142
	Total	169	126	295

Based on Table 6.1, it was found that the automatic speech intelligibility detection commits more errors using SVM where not-intelligible is more frequently misclassified as intelligible 6 times (Type I error) and intelligible is misclassified as not-intelligible 18 times (Type II error).

Table 6.2 shows the classification accuracy of the proposed FPN for all 10 folds. The average of training and testing accuracy are also calculated as the performance of the classifier.

Table 6.2: Classification accuracy of the proposed FPN for all folds

Fold	FPN	
	training	testing
Fold 1	98.76	99.32
Fold 2	99.17	99.66
Fold 3	98.87	97.63
Fold 4	99.10	97.96
Fold 5	99.02	98.31
Fold 6	99.06	98.98
Fold 7	98.95	97.63
Fold 8	98.98	99.32
Fold 9	98.76	97.30
Fold 10	98.98	96.95
Average	98.96	98.31

Table 6.3 presents the classification accuracy, precision and recall of the proposed FPN in terms of the training and testing. The classification accuracy is 98.31%, precision is 100.00% and recall is 94.40%.

Table 6.3: Classification accuracy, precision and recall of the proposed FPN

Accuracy	98.31%
Precision	100.00%
Recall	94.40%

Table 6.4 shows the classification accuracy of FPN for individual speech features and the mean values of the aspect of speech which are prosody, pronunciation and recall. Overall, 3rd MFCC coefficient produces the highest accuracy with 86.14%. Meanwhile, the lowest accuracy is the 6th MFCC coefficient with 60.89%. The mean values for prosody is 72.11%, pronunciation is 75.56% and voice quality is 81.48%.

Table 6.4: The classification accuracy based on the individual speech features

Aspect of speech	Speech features	Evaluation
Prosody	F0	72.62
	Energy	62.42
	ZCR	81.29
	Mean	72.11
Pronunciation	MFCC 0	81.29
	MFCC 1	61.74
	MFCC 2	83.57
	MFCC 3	86.14
	MFCC 4	75.30
	MFCC 5	79.91

	MFCC 6	60.89
	MFCC 7	74.99
	MFCC 8	69.81
	MFCC 9	81.36
	MFCC 10	72.08
	MFCC 11	81.36
	MFCC 12	73.81
	Mean	75.56
Voice quality	Jitter	79.87
	Shimmer	83.09
	Mean	81.48

6.6 Discussion of Findings

Based on the classification results, discussion of the findings are as follows;

6.6.1 Accuracy, Precision and Recall: FPN vs Baseline Methods

Figure 6.9 presents the comparisons of the classification accuracy, precision and recall of the proposed FPN with the benchmark classifiers. FPN has achieved a promising results with the highest classification accuracy (98.31%), and precision (100%) among all the classifiers. For recall, FPN produces 94.42% which is the second highest after RF (95.67%).

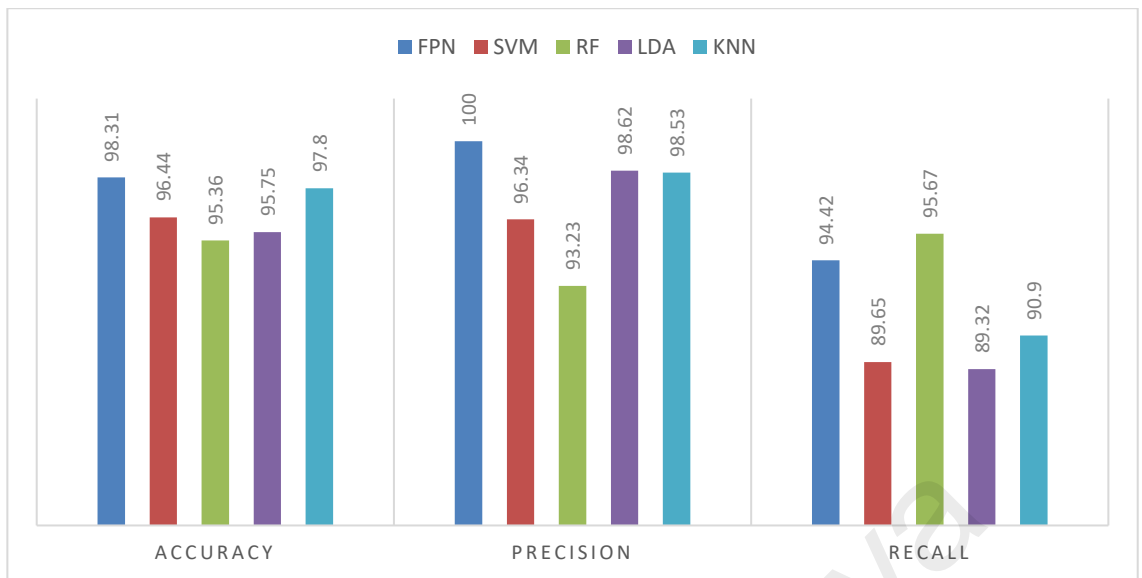


Figure 6.9: Comparisons of the classification accuracy, precision and recall of FPN with the benchmark classifiers

Table 6.5 compares the mean values of accuracy, precision and recall for FPN with baseline classification methods. In terms of the mean value for accuracy, precision and recall, FPN produces the highest mean value with 97.57%. This indicates that FPN has greater discrimination ability than the baseline classification methods in detecting the intelligibility of impaired speech.

Table 6.5: The mean values of accuracy, precision and recall for FPN and baseline methods

Classification method	Accuracy	Precision	Recall	Mean value
FPN	98.31	100.00	94.40	97.57
SVM	96.44	96.34	89.65	94.14
RF	95.36	93.23	95.67	94.75
LDA	95.75	98.62	89.32	94.56
KNN	97.80	98.53	90.90	95.74

6.6.2 Misclassification and Error Rate

For analysis of the type of errors, false positive is known as Type I error and false negative is known as Type II error. Table 6.6 compares the Type I and Type II errors for FPN and the baseline methods in terms of the frequencies with the false positive rate intelligibility ($p(\text{FP}_i)$) and false negative rate of intelligibility ($p(\text{FN}_i)$).

Table 6.6: Comparison of classification errors

Classification method	Type I		Type II	
	Frequency	$p(\text{FP}_i)$	Frequency	$p(\text{FN}_i)$
FPN	0	0.00	9	0.06
SVM	6	0.05	18	0.11
RF	10	0.07	7	0.05
LDA	2	0.02	18	0.11
KNN	3	0.02	15	0.09

For type I error, RF is the highest with 10 times of incorrect rejection of *not intelligible*. For Type II error, SVM and LDA generates 18 times of failure to reject the false *intelligible*. On the other hand, FPN has no Type I error and Type II 9 times. For Type I rate, FPN has the lowest with 0.00 rate and RF with the highest at 0.07. For Type II rate, the lowest is RF at 0.05. SVM and LDA, both have the highest rate at 0.11.

6.6.3 The Classification Accuracy of Speech Features: FPN vs Baseline Methods

Table 6.7 show the classification accuracy for each individual speech features for FPN and baseline methods. The mean values for each aspect of speech is also presented.

For prosody, RF produces the highest accuracy at 74.55% and SVM the lowest at 63.01%. For pronunciation, FPN produces the highest accuracy at 75.56% and KNN the lowest (67.17%). For voice quality, FPN has the highest accuracy at 81.48% and KNN the lowest (75.89%).

Table 6.7: The classification accuracy of FPN and baseline classifiers for individual speech features

Aspect of speech	Speech features	FPN	SVM	RF	LDA	KNN
Prosody	F0	72.62	56.73	70.08	61.16	67.23
	Energy	62.42	51.71	74.69	51.81	56.35
	ZCR	81.29	80.58	78.89	80.58	78.55
	Mean	72.11	63.01	74.55	64.52	67.38
Pronunciation	MFCC 0	81.29	80.89	77.33	80.92	73.77
	MFCC 1	61.74	61.60	57.30	61.74	50.53
	MFCC 2	83.57	83.19	79.67	83.19	78.42
	MFCC 3	86.14	86.04	84.34	86.17	74.72
	MFCC 4	75.30	73.16	71.16	72.86	71.06
	MFCC 5	79.91	79.97	77.40	80.07	75.84
	MFCC 6	60.89	59.57	56.39	59.03	55.30
	MFCC 7	74.99	74.96	68.62	74.96	66.96
	MFCC 8	69.81	69.57	63.94	69.37	61.71
	MFCC 9	81.36	81.23	77.60	81.23	75.23
	MFCC 10	72.08	72.14	66.15	71.87	64.59
	MFCC 11	81.36	73.43	69.61	73.37	66.66
	MFCC 12	73.81	73.20	71.60	73.13	58.39
	Mean	75.56	74.53	70.85	74.45	67.17
Voice quality	Jitter	79.87	73.50	75.09	76.38	74.21
	Shimmer	83.09	78.99	79.02	82.04	77.57

	Mean	81.48	76.25	77.06	79.21	75.89
--	------	-------	-------	-------	-------	-------

6.6.4 Best Speech Features in Detecting Speech Intelligibility

Table 6.8 shows the mean values for prosody, pronunciation and voice quality aspect of all classification methods. Voice quality produces the highest mean at 77.98%, followed by pronunciation (72.51%) and prosody (68.31%).

Table 6.8: The mean values for prosody, pronunciation and voice quality

	FPN	SVM	RF	LDA	KNN	Mean
Prosody	72.11	63.01	74.55	64.52	67.38	68.31
Pronunciation	75.56	74.53	70.85	74.45	67.17	72.51
Voice quality	81.48	76.25	77.06	79.21	75.89	77.98

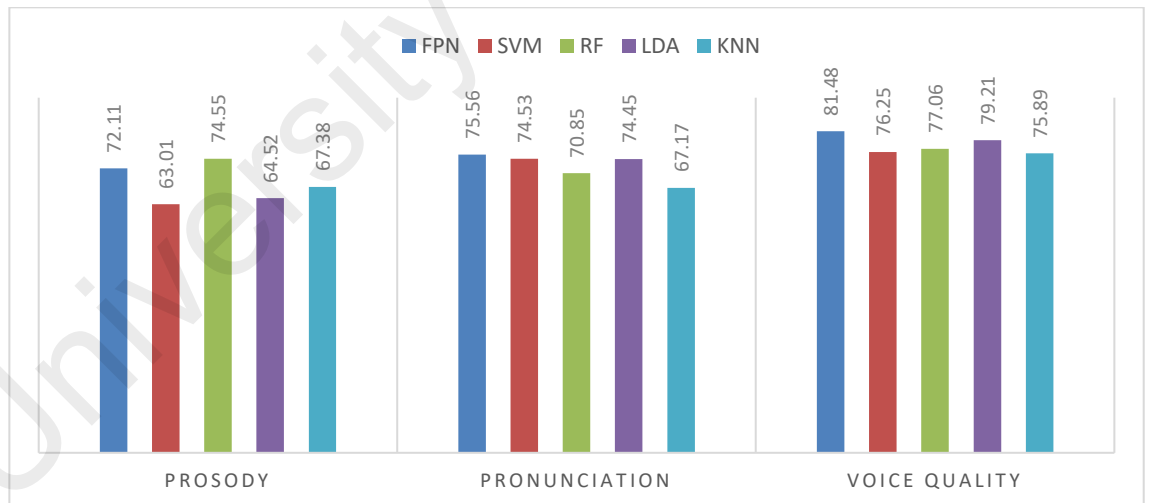


Figure 6.10: The mean values of prosody, pronunciation and voice quality for FPN and baseline methods

Based on the comparison graph in Figure 6.10, voice quality indicates the highest accuracy for all classification methods. This indicates that jitter and shimmer are significant speech features for detecting speech intelligibility of impaired speakers.

In addition, a linear regression analysis is performed to determine the effect of prosody, pronunciation and voice quality to intelligibility detection. The purpose is to understand which of the speech aspect is statistically significant predictor for intelligibility detection in impaired speech. Figure 6.11 shows the effect of prosody, pronunciation and voice quality on classification accuracy.

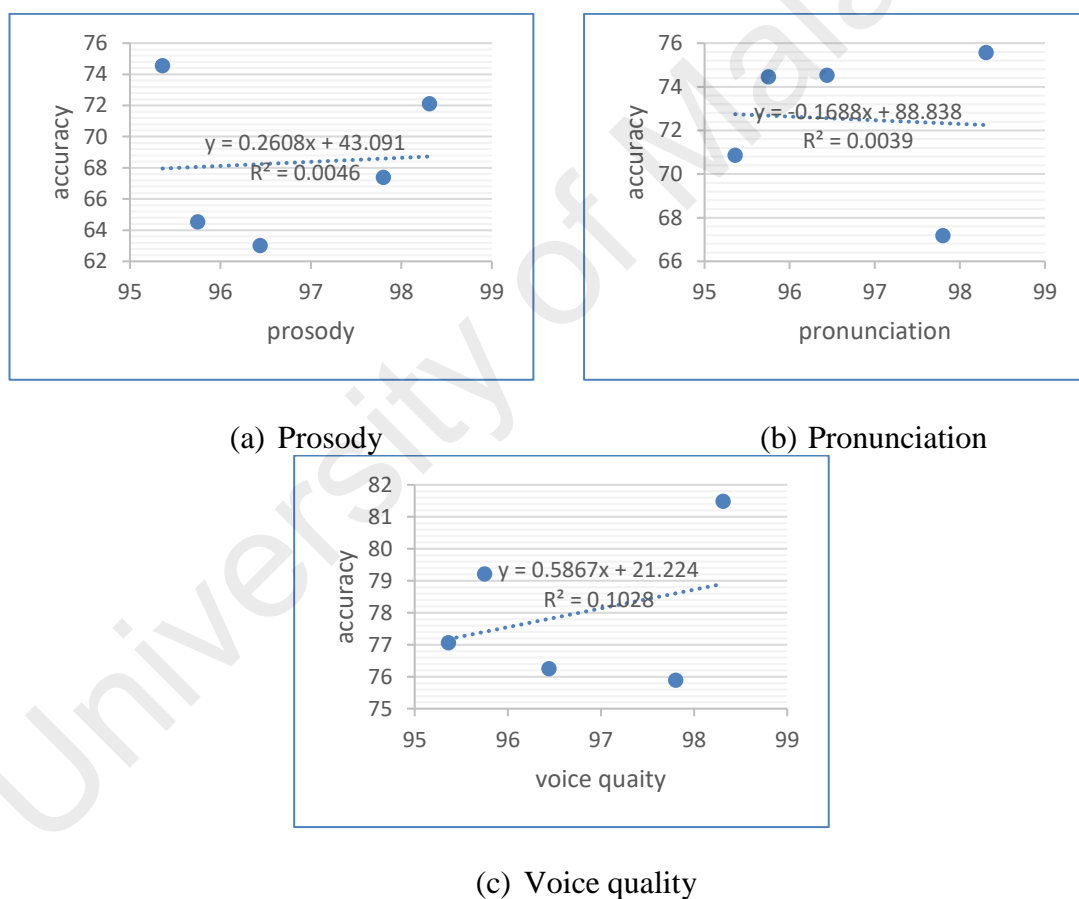


Figure 6.11: Effect of (a) prosody, (b) pronunciation and (c) voice quality on classification accuracy

Pronunciation and voice quality are found to be a significant predictors of classification accuracy, ($p=0.008$, >0.005) and ($p=0.038$, >0.005), respectively. However, prosody is found to be insignificant. The variance in classification accuracy can be explained by prosody (0.5%), pronunciation (0.4%) and voice quality (10.3%). Among the three, voice quality is found to explain more on the variation in classification accuracy. Table 6.9 summarizes the correlation and coefficient of determination of prosody, pronunciation and voice quality.

Table 6.9: Correlation and coefficient of determination of prosody, pronunciation and voice quality

	R	R ²	F	P
Prosody	0.068	0.005	0.014	0.003
Pronunciation	0.062	0.004	0.012	0.008
Voice quality	0.321	0.103	0.344	0.038

As discussed in Section 4.9, voice quality such as jitter and shimmer are the significant aspect of speech for impaired speech that causes the intelligibility deficits. These two features correlates with the hoarseness in speech, which reduce the quality of speech for impaired speakers (Vipperla, 2010). This is because, speech impaired children have speech abnormality that affects the vocal folds, either muscle or neural activity involved with phonation, either lesions that may cause increase in aperiodicity of vocal fold vibration which was reflected in the increased value of jitter (Wertzner et al., 2005). The speech characteristics is also indicated by the reduction of glottic resistance, vocal fold mass lesions and greater noise at production, which are some of the factors that influence shimmer values (Wertzner et al., 2005). Therefore, in this research, we have identified that voice quality that consists of jitter and shimmer have

more discriminative power in detecting speech intelligibility of impaired speech compared to prosody and pronunciation aspect.

6.7 Comparison with Existing Work

Table 6.10 presents the existing research in FPN that has been used in various application of detection system. The comparison of FPN approach in detection system such as knowledge representation and inference engine as well as tools used for implementation is provided.

University of Malaya

Table 6.10: Summary of related work on fuzzy petri nets for detection system

Research	Application	Detection				Implementation
		Input	Output	Inference engine		
				Knowledge representation	Inference mechanism	
Author (2016)	Speech intelligibility detection	Salient discriminative speech features in intelligibility such as; F0, energy, ZCR, MFCC 0 th – 12 th coefficient, jitter, shimmer	Speech intelligibility	FPN	FIS	MATLAB PN Editor GPenSIM
Ivasic-Kos et. al (2014) Image classification	Image detection	16 image features based on colour, position, size and shape of the region	Multi-level image classes from four semantic levels as follows - An elementary class - A generalization class - A derived class - A scene class	KRFPN	-Fuzzy inheritance, -Fuzzy recognition -Fuzzy intersection	Corel image dataset Normalized Cut algorithm
Szwed (2014)	Video event detection	Video sequences and object	Video event	Fuzzy Semantic Petri Nets	-Fuzzy ontology -Fuzzy description	JAVA

					logic	
Shen et al (2013)	Human fall detection	Human body inclination and the occurrence frequencies at the peak of the area of use	Human body condition: -Walking -Exercising -Falling down	High-Level FPN	Fuzzy rules	Not available
Kouzehgar et. al (2011)	Human Behavior Verification and Validation	Questionnaire	Probable structural and semantic errors of human behaviour	FPN	Fuzzy rules	Not available
Cheng and Yang (2009)	Railway traffic control	-Possible dispatching decision factors -Possible dispatching options	Abnormal event	FPN	Fuzzy rules	Not available
Lee et.al (1999)	Damage assessment of bridges	All information is shown hierarchically from overall damage level to inspection details such as follows <ul style="list-style-type: none"> • Damage level (A Fuzzy degree + truth values (TV)) 	Visual inspection	Fuzzy rules + Truth Value (TV)	F -> G, TV1 F1 TV2 G1, TV3	JAVA

		<ul style="list-style-type: none">• Damage cause (A cause + TV)• Recommendations				
--	--	---	--	--	--	--

University of Malaya

6.8 Summary

This chapter presents the development of the proposed FPN for speech intelligibility detector. From the experiments, several important findings have been identified as follows:

- The proposed FPN outperforms the baseline classification methods in classification accuracy at 98.31%, and recall at 100.00%.
- In detection of individual speech features, FPN outperforms the baseline classification methods with mean value of 81.48%.
- For misclassification rate, FPN produces the lowest Type I error of 0.00%. In Type II error, FPN produces the second lowest rate (0.06) after RF (0.05).
- Among the three aspects of speech, voice quality has been identified as a significant speech features for detecting speech intelligibility for children with speech impairment.

CHAPTER 7 CONCLUSION

This chapter discusses the overall work carried out in this research. First, the research objectives identified in chapter 1 are revisited. Second, research contributions is presented. Third, some limitations of this research is explained. Finally, recommendations to the future work are provided.

7.1 Research objectives revisited

This section revisits the achievements of the research objectives identified in this research.

7.1.1 Research objective 1

The first objective is to identify the significant impaired speech features that are significant for the performance of automatic speech intelligibility detection. To achieve this objective, we begin with the analysis of literature in Chapter 2, accumulating and analyzing the speech corpus in Chapter 4 and the experimental work performed in Chapter 5. For objective 1, there are two research questions that need to be considered as follows:

RQ1: What are relevant speech features that could potentially affect the intelligibility of impaired speech?

RQ2: What speech features should be measured in Malay pronunciation of speech impaired speakers in automatic speech intelligibility detection?

In answering RQ1, the relevant speech features that could potentially affect the intelligibility of impaired speech is discussed in Chapter 2. According to Kim et al (2015), these features can be categorized in prosody, pronunciation and voice quality.

The selection of speech features is reflected to the types of speech impairments presented in Chapter 1 which are articulation disorders, voice disorders and fluency disorders. Figure 7.1 show the mapping of the types of speech impairments to the category or aspect of speech features.

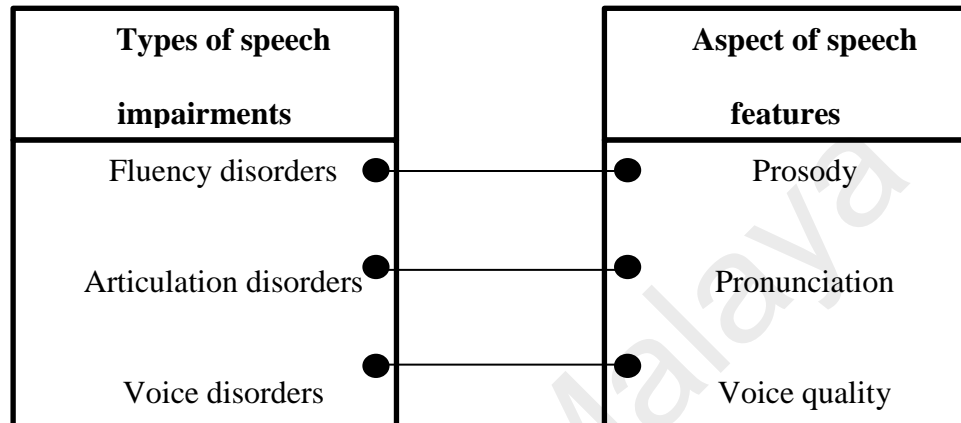


Figure 7.1 Mapping of the types of speech impairments and the aspect of speech features

The three types of disorders are influenced by the speech features of impaired speech. For fluency disorders, the speaking flow is interrupted, which affect the atypical rate, rhythm and repetition in sounds. The speech features related to prosody such fundamental frequency (F0) or pitch, intensity, energy and normalized duration of syllables are related to tone, loudness and rhythm structures. These features are suitable representation of the characteristics of impaired speech with fluency disorders. Articulation disorders are correlated with the ability of the articulator to pronounce words. Therefore, speech features that carry meaningful information related to pronunciation such as MFCC is important to represent the characteristics of articulation disorders. For voice disorders, this is related to the abnormal production in voice quality which includes aspect of phonation and resonance. Therefore, speech features related to voice quality are important to represent the characteristics of voice disorders.

For RQ2, the speech features that should be measured for Malay pronunciation of impaired speakers in automatic speech intelligibility detection have been identified in chapter 3 and 4. Relevant speech features in Malay impaired speech are F0, energy, ZCR, MFCC, jitter and shimmer. In chapter 5 and 6, we have identified that the most significant speech features of Malay impaired children are related to jitter and shimmer.

7.1.2 Research objective 2

The second objective is to identify suitable classification method to enhance the performance of automatic speech intelligibility detection for speech impaired speakers. The identification is carried out with information gained from the analysis of literature and justification in Chapter 2 and 3. Research questions related to objective 2 are as follows;

RQ3: What is needed to optimize the discrimination ability in the automatic speech intelligibility detection of impaired speech?

RQ4: What are the basis, framework and contents of the identified classification method?

The identification of suitable classification method is based on the lacking of the discrimination ability on the existing methods. Discrimination ability in detection system makes use of the speech features and the classification methods itself. The main issue is the characteristics of impaired speech, which is difficult to be classified due to high variability and confusability that lead to low intelligibility of speech. In answering RQ3, it is important to have a classifier that has the ability to reason the speech knowledge to decide the particular class of intelligibility. For that, FPN is identified as a suitable method that has the greater knowledge representation ability to reason ambiguous information.

In RQ4, the basis, framework and contents of the identified classification methods explained in Chapter 3.

7.1.3 Research objective 3

The third objective is to develop an automatic speech intelligibility detector based on the identified classification methods in objective 2. In this stage, we formulate a framework that can support the development of Fuzzy Petri Nets for speech intelligibility detector. This is explained in Chapter 3 and 5. One research question for objective 3 is as follows:

RQ5: How the proposed method developed?

The proposed method development is explained in Chapter 6.

7.1.4 Research objective 4

The fourth objective is to evaluate and compare the performance of the proposed classification method with the existing benchmark methods. The system prototype is developed which is presented in Chapter 6. The benchmark methods are presented in Chapter 5. Research questions related to objective 4 are as follows:

RQ6: What are the measurement used to evaluate the proposed methods?

RQ7: How the results of the proposed method being compared to the baseline method?

For RQ6, there are several measurements used to evaluate the proposed methods as explained in Section 3.7. Table 7.1 recaps the types of evaluation used to evaluate the proposed FPN and their purposes.

Table 7.1: Types of evaluation and their purposes

Types of evaluation	Purposes
The misclassification rate	Type I – To measure the false positive rate for a given class Type II – To measure the false negative rate a given class
Classification accuracy	To calculated the percentage of the correct prediction
Precision	To calculate the percentage of the correct positive prediction
Recall	To calculate the percentage of the positive cases

In answering RQ7, the comparison of the results for proposed FPN and the baseline methods are discussed in Section 6.6. The comparisons are made to identify the performance of the proposed FPN and the baseline methods using the identified evaluation measurements. In addition, comparison is also made to understand significant speech features with high discrimination ability in detecting the speech intelligibility of impaired speech among Malay speaking children.

7.2 Research Contribution

This research contribution includes the following;

A corpus of Malay children speakers with speech impairments

It is identified that the difficulty of the access to suitable corpus Malay children speakers with speech impairments. Therefore, in this research, a corpus of Malay

children speakers with speech impairments is developed. The corpus consist of 30 impaired speakers with 60 control speakers. The development and analysis of the Malay speech corpus is in achieving objective 1.

Speech analysis of significant speech features of Malay children speakers with speech impairments in intelligibility detection

The speech analysis performed in achieving objective 1 is to identify the significant speech features for Malay speaking children with speech impairments for the automatic speech intelligibility detection. It is important to identify the significant speech features that has high discriminative ability between impaired and control group. The selection of relevant speech features contribute to the performance of the detector. Based on the speech analysis, we have identified relevant speech features for Malay speaking children with speech impairments such as F0, energy, ZCR, MFCC, jitter and shimmer. The significant features among all for detecting speech intelligibility are jitter and shimmer. The discussion of findings is explained in Section 6.6.4

Improved automatic speech intelligibility detector for children with speech impairments

The proposed FPN has shown the success of improving the discrimination ability of the automatic speech intelligibility detector for children with speech impairments. The justification and implementation of the proposed approach is presented in Chapter 3 and 6.

Comparative results of the proposed classifier with existing baseline methods

The evaluation of the proposed FPN and the existing baseline methods covers misclassification rate, accuracy, precision and recall. The proposed classifier has outperformed all the baseline classifiers for all the above performance measures.

7.3 Research Limitation

Speech resource limitation. This research suffers from resource limitation that could potentially influence the outcome of the research. As the focus is on children, very limited data was collected due to difficulties in obtaining and handling recording sessions with speech impaired children. However, the focus of this research is more on classification, thus this limitation may not be that significant.

Human expert. The machine learning techniques perform classification based on a set of input by human expert, which may contain errors. Although this research has taken the necessary due care, it is not guaranteed that all human errors has been eliminated.

Knowledge on the FPN approach and simulation. Though we have identified the suitability and performance of FPN in detecting speech intelligibility of impaired speech, this research has limitation. The proposed method presented in this thesis requires medium to advanced knowledge in the in FPN technology and simulation tool to perform the simulation. It is not the same as the other methods such as SVM, KNN, RF and LDA which their toolboxes are already provided in MATLAB.

7.4 Future work

Though FPN has been proven to increase the ability of detector and increase the system performance, there are still rooms for improvements. This section provides several suggestions which may be consider to extend the research in terms of short-term and long-term goals.

Short-term goals

Considering more features. This research has been focusing on the prosody, voice quality and pronunciation aspect of speech to improve the classification performance. Further research can be conducted in using other combinations of speech features such as Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP) and harmonic to noise ratio (HNR) in order to maximize the classification performance.

Optimizing the learning ability in FPN. FPN is proven as a classification method that has the capability in classifying the speech intelligibility of impaired speech. It is a powerful modeling tool for fuzzy production rules-based knowledge systems. The classification performance can also be improved by optimizing the learning ability of FPN. It is interesting to see how this implicates the study.

Long-term goals

Enhance the development of the back end of DBASR. This research only focuses on the front end of the DBASR which involves the classification task. For further research, this system can be enhance to the back end of DBASR to develop a complete DBASR. Further works can be focus on the linguistic merger and the recognition process of the DBASR.

Development of the applications based Automatic Speech Intelligibility Detection. For further research work, many useful applications based on the automatic speech intelligibility detection can be developed. Example of the applications such as the intelligibility assessment tool for diagnostic and therapy evaluation, capturing the abnormal speech variation tools as well as opportune treatment in order to assist speech therapist and SLPs. On the other hand, another applications which are useful for the

speech impaired speakers, caretakers and teachers can be developed. Example of the applications are the assistive and educational technology such as the personal speech tracker and assessment.

University of Malaya

REFERENCES

- American Speech and Hearing Association (ASHA). Dysarthria. Retrieved from <http://www.asha.org/public/speech/disorders/dysarthria.htm>.
- Aggarwal, J. K. and Cai, Q. (1999). Human motion analysis. *Journal Computer Vision and Image Understanding*, 73(3): 428-440.
- Al-Haddad, S., S. Samad, (2008). Isolated Malay digit recognition using pattern recognition fusion of dynamic time warping and hidden Markov models. *American Journal of Applied Sciences*, 5(6): 714-720.
- Amano-Kusumoto, A., J.-P. Hosom, Kain, A., and Aronoff, J.M. (2014). Determining the relevance of different aspects of formant contours to intelligibility. *Journal Speech Communication*, 59: 1-9.
- An, A. (2005). Classification methods. Retrieved from <http://www.cs.yorku.ca/~aan/research/paper/238An.pdf>
- Anumanchipalli, G. K., Meinedo, H. and Bugalho, M. (2012). *Text-dependent pathological voice detection*. Proceeding of Interspeech. Portland, Oregon, U. S. A: 530–533.
- Anusuya, M. A. and S. K. Katti (2009). Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, 6(3): 181-205.
- Ashraf, J., Iqbal, N., Khattak N. S. and Zaidi, A. M. (2010). Speaker Independent Urdu speech recognition using HMM. *Informatics and Systems*, 1-5.
- Attieh, A. A (2010). Linguistic Factors Affecting the Loci and frequency of Stuttering across Age Groups in Arabic-Speaking Jordanians. *Journal of the Royal Medical Services*, 17(3).
- Bälter, O., Engwall, O., Öster, A-M. and Kjellström, H. (2005). *Wizard-of-Oz Test of ARTUR – A Computer-Based Speech Training System with Articulation Correction*. In Proceedings of the Conference on Computers and Accessibility, Baltimore.
- Bauman-Waengler, J. (2012). *Articulatory and Phonological Impairments: A Clinical Focus*, Allyn & Bacon Communication Sciences and Disorders Series, Pearson, New Jersey
- Berenthal, J., Bankson, N. and Flipsen, P. (2009). *Speech Sounds Disorders*. Boston, Allyn & Bacon.
- Blaney, B. and Wilson, J. (2000). Acoustic variability in dysarthria and computer speech recognition. *Clinical Linguistic and Phonetic* 14(4): 307-327.

- Boersma, P. and Weenink, D. Praat: doing phonetics by computer. Retrieved from <http://www.fon.hum.uva.nl/praat/>
- Boswell, D. (2002). Introduction to Support Vector Machines. Retrieved from <http://dustwell.com/PastWork/IntroToSVM.pdf>
- Brand, M. and V. Kettner (2000). Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 844–851.
- Breiman, L. (2001) Random Forests. *Machine Learning*, 45(1), pp 5-32
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees, Wadsworth International Group
- Brockmann, B. (2011). *Improving jitter and shimmer measurements in normal voices*. Ph.D, Thesis, Medical School Institute of Cellular Medicine
- Bromberg, I., Fu, Q., Hou, J., Li, J., Ma, C., Matthews, B., Morena-Daniel, A., Morris, J., Siniscalchi, S. M., Tsao, Y. and Wang, Y. (2007). *Detection-Based ASR in the Automatic Speech Attribute Transcription Project*. INTERSPEECH. Antwerp, Belgium: 1829 - 1832.
- Butt, A. H. (2012). *Speech assessment for the classification of hypokinetic dysarthria in Parkinson's disease*. Masters Dissertation, Computer Engineering, Dalarna University
- Campbell, W. M., Assaleh, K. T. and Brown, C. C. (1999). Low-complexity small-vocabulary speech recognition for portable devices. *Signal Processing and Its Applications*, 6: 619-622.
- Canterla, A. M. d. H. (2012). *Design of Detectors for Automatic Speech Recognition*. Ph.D Thesis, Department of Electronics and Telecommunications, Norwegian University of Science and Technology.
- Chen, X., and Jin, D. (2002). *Fuzzy Petri Nets for Rule-based Pattern Classification*. IEEE International Conference on Communications, Circuits and Systems and West Sino Expositions, 2: 1218 - 1222
- Cheng, Y.-H. and Yang, L.-A. (2008). A Fuzzy Petri Nets approach for railway traffic control in case of abnormality: Evidence from Taiwan railway system. *Expert system with Applications*, 36(4): 8040-8048
- Choi, D.-L., and Kim, B.-W. (2011). *Dysarthric Speech Database for Development of QoLT Software Technology*. Eight International Conference on Language Resources and Evaluation, Istanbul, Turkey.
- Cohn, J. F., and Kruez, T. S. (2009). *Detecting Depression from Facial Actions and Vocal Prosody*. 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. Amsterdam.

- Coleman, C. and Meyers, L. (1991). Computer recognition of the speech of adults with cerebral palsy and dysarthria. *Augmentative Alternative Communication*, 7(1): 34–42.
- Colton, R. H. and Casper, J. K. (2006). Understanding voice problems: A physiological perspective for diagnosis and treatment. Baltimore, Lippincott Williams & Wilkins.
- Connolly, J. H. (1986). Intelligibility: a linguistic view. *International Journal of Language & Communication Disorders*, 21(3): 371–376.
- Cutler, A., Dahan, D. and Donselaar, W. v. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, vol. 40, pp. 141-201, 1997
- Darley, F., Aronson, A., and Brown, J. R. (1969). Differential diagnostic patterns of dysarthria. *Journal of speech, language and hearing research*, 12: 246-269.
- Davidrajuh, R. (2007). *Design and Application of Templates in GPenSIM*. Pacific Asia Conference on Information Systems (PACIS).
- Davis, S. B. (1978). Acoustic characteristics of normal and pathological voices. Report Haskin Laboratories.
- Deller, J. R., and Hsu, D. (1991). On the use of the Hidden Markov Model modelling for recognition of dysarthric speech. *Computer methods and Programs in Biomedicine*, 35: 125-139.
- Doyle, P. C., Leeper, H. A., Kotler, A. L., Thomas-Stonell, N., O'Neill, C., Dylke, M. C., and Rolls, K. (1997). Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility. *Journal of Rehabilitation Research and Development*, 34(3): 309-316.
- Eguchi, S. and Hirsh, I. J. (1969). Development of speech sounds in children. *Acta Oto-Laryngologica*, vol. Supplementum 257: 1–51.
- El-Imam, Y. A. and Don, Z. M. (2005). Rules and Algorithms for Phonetic Transcription of Standard Malay. *IEICE TRANSACTIONS on Information and Systems*, E88-D(10): 2354-2372.
- Farrús, M. (2007). *Jitter and shimmer measurements for speaker recognition*. Proceedings of the International Conference Interspeech 2007.
- Feng, L., and Obayashi, M. (2012). Construction and Application of Learning Petri Net. *Petri Nets - Manufacturing and Computer Science*. P. Pawlewski, intech.
- Ferrier, L., Shane, H., Ballard, H., Carpenter, T., and Benoit, A. (1995). Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11(3): 165-175.

- Fook, C. Y., and Muthusamy, H. (2013). Comparison of speech parameterization techniques for the classification of speech disfluencies. *Turkish Journal of Electrical Engineering & Computer Sciences*.
- Fougeron, C., Crevier-Buchman, L., Fredouille, C., Ghio, A., Meunier, C., Chevrie-Muller, C., Audibert, N., Bonastre, J.-F., Colazo Simon, A., Delooze, C., Duez, D., Gendrot, C., Legou, T., Levèque, N., Pillot-Loiseau, C., Pinto, S., Pouchoulin, G., Robert, D., Vaissiere, J., Viallet, F., and Vincen, C. (2010). The DesPho-APaDy Project: Developing an Acoustic-phonetic Characterization of Dysarthric Speech in French. *International Conference on Language Resources and Evaluation, Valletta, Malta*.
- Freitas, A.A., Wieser, D. C., and Apweiler R. (2010). On the Importance of Comprehensible Classification Models for Protein Function Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1): 172-182
- Fudala, J. B. (2000). Arizona Articulation Proficiency Scale, Arizona TM.
- Giuliani, D., M. Gerosa. (2006). Improved automatic speech recognition through speaker normalization. *Computer Speech and Language*.
- Goldstein, U. G. (1980). *An Articulatory Model for the Vocal Tracts of Growing Children*. Ph.D. Thesis, Cambridge, Massachusetts Institute of Technology.
- Gorelick, L., M. Blank, (2007). Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12): 2247–2253.
- Green, R. (1966). Linguistic subgrouping within Polynesia: The implications for prehistoric settlement. *Journal of the Polynesian Society*, 75: 6-38.
- Hacioglu, K., Chen, Y. (2005). *Automatic Time Expression Labeling for English and Chinese Text*. 6th International Conference Computational Linguistics and Intelligent Text Processing: 548–559.
- Harrington, J. (2010). *The Phonetic Analysis of Speech Corpora*, Wiley-Blackwell.
- Harrington, J., Palethorpe, S., Watson, C. (2005). Deepening or lessening the divide between diphthongs: an analysis of the Queen's annual Christmas broadcasts. New Jersey, Lawrence Erlbaum.
- Hartl, D. A., Hans, S., Vaissière, J., and Brasnu, D. A. (2003). Objective acoustic and aerodynamic measures of breathiness in paralytic dysphonia. *Eur Arch Otorhinolaryngology*, 260(4): 175-182.
- Haynes, W. O. and R. H. Pindzola (2012). *Motor Speech Disorders, Dysphagia, and the Oral Exam*, Pearson Education Inc.
- Hellmann, M. (2001) Fuzzy Logic Introduction. Retrieved from <http://www.ece.uic.edu/~cpress/ref/2001Hellmann%20fuzzyLogic%20Introduction.pdf>

- Hosom, J. P., Kain, A. B., Mishra, T., van Santen, J. P. H., Fried-Oken, M., and Staehely, J. (2003). *Intelligibility of modifications to dysarthric speech*. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong.
- Huang, D.-Y., and Dong, M. (2014). *Intelligibility detection of pathological speech using asymmetric sparse kernel partial least squares classifier*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence.
- Huang, X., Acero, A., and Hon, H-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall.
- Hunt, A. (1997). *Comp.Speech; Frequently Asked Question*, Speech Applications Group, Sun Microsystems Laboratories.
- Hux, K., Rankin-Erickson, J., Manasse, N., and Lauritzen, E. (2000). Accuracy of three speech recognition systems: Case study of dysarthric speech. *Augmentative Alternative Communication*, 16(3): 186–196.
- Ikui, Y., Tsukuda, M., Kuroiwa, Y., Koyano, S., Hirose, H., Taguchi, T. (2011). Acoustic characteristics of ataxic speech in Japanese patients with Spinocerebellar Degeneration. *International Journal of Language Communication Disorders* 47(1): 84-94.
- Ivasic-Kos, M., Ribaric, S. and Ipsic, I. (2014). Multi-level Image Classification Using Fuzzy Petri Net. *Recent Advances in Neural Networks and Fuzzy Systems*.
- Jayaram, G. and Abdelhamied, K. (1995). Experiments in dysarthric speech recognition using artificial neural networks. *Journal of Rehabilitation Research and Development*, 32(2): 162–169.
- Jeng, J. Y., Weismer, G. and Kent, R. D. (2006). Production and perception of mandarin tone in adults with cerebral palsy. *Clinical Linguistics & Phonetics*, 20(1): 67-87.
- Jurafsky, D. and J. H. Martin (2008). *Speech and Language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA, Prentice Hall.
- Kain, A. B., Hosom, J.-P. Niu, X., van Santen, J. P. H., Fried-Oen, M., and Staehely, J. (2007). Improving the Intelligibility of Dysarthric Speech. *Speech communication* 49(9): 743–759.
- Kates, J. M. (2008). *Digital Hearing Aids*. San Diego, California, USA, Plural Publishing.
- Kent, R. D. (1976). Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of Speech and Hearing Research*, 19: 421–447.

- Kent, R. D., J. Kent, Duffy, J. R., Thomas, J. E., Weismer, G. and Stuntebeck, S. (2000). Ataxic dysarthria. *Journal of Speech, Language, and Hearing Research*, 43(5): 1275-1289.
- Kent, R. D., Kent, J. F., Duffy, J. R., and Weismer, G. (1998). The Dysarthrias: speech voice profiles, related dysfunction, and neuropathology. *Journal of Medical Speech Language Pathology*, 6(4): 165-211.
- Kent, R. D., Weismer, G., Kent, J. F., and Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders* 54: 482-499.
- Khan, T., Westin, J., and Dougherty, M. (2014). Classification of Speech Intelligibility in Parkinson's Disease. *Biocybernetics and Biomedical Engineering*, 34: 35-45.
- Kim, H., Drake, B. L., and Park, H. (2007). Multiclass classifiers based on dimension reduction with generalized LDA. *Pattern Recognition*, 40(11): 2939-2945
- Kim, J., Kumar, N., Tsiartas, A., Li, M., and Narayanan, S. S. (2015). Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech & Language*.
- Kouzehgar, M., Badamchizadeh, M. A., Khanmohammadi, S. (2011). *Fuzzy Petri Nets for Human Behavior Verification and Validation*. International Journal of Advanced Computer Science and Applications, (IJACSA) 2(12): 106-114.
- Ladefoged, P. and I. Maddieson (1996). *The Sounds of the World's Languages*. Oxford, Blackwell.
- Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). *Learning realistic human actions from movies*. IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK.
- Lapteva, O. (2011). *Speaker Perception and Recognition: An Integrative Framework for Computational Speech Processing*, Kassel University Press.
- Latifov, A. (2012). *Dynamic Enterprise Architecture - From Static to Dynamic Models*. Computer science. University of Stavanger, University of Stavanger: 78.
- Lee, J., Liu, K. F. R., and Chiang, W. (1999). A Fuzzy Petri Net-Based Expert System and Its Application to Damage Assessment of Bridges. *IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics*, 29: 350-369.
- Lee, S., Potamianos, A., Narayanan, S. S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*: 1455–1468.
- Lemmetty, S. (1999). *Review of Speech Synthesis Technology* Helsinki University of Technology.
- Li, J. and Lee, C.-H. (2005). *On designing and evaluating speech event detectors*

Interspeech, 2005, Lisboa.

- Li, J., Tsao, Y. and Lee, C. H. (2005). *A Study on Knowledge Source Integration for Candidate Rescoring in Automatic Speech Recognition*. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). pp. 837 - 840
- Liss, J. M., Spitzer, S. M., Caviness, J. N., and Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *Journal of Acoustical Society of America* 112: 3022-3030.
- Liu, B. (2011). *Simulation Network Intrusion Detection System with GPenSim*. Faculty of Science and Technology. Stavanger, Norway, University of Stavanger. Computer Science: 42.
- Liu, X. and Yin, G.-S. (2009). *Fuzzy Neural Petri Nets for Expert Systems*. Intelligent Computation Technology and Automation, 2009. ICICTA '09. Changsha, Hunan: 732-735.
- Martland, P., Whiteside, S. P., Beet, S. W., and Baghai-Ravary, L. (1996). *Estimating child and adolescent formant frequency values from adult data*. International Conference Speech Language Processing. Philadelphia: 626–630.
- Meher Taj, S. and Kumaravel, A. (2015). Survey on Fuzzy Petri Nets for classification. *Indian Journal of Science and Technology*, 8(40): 1-8
- Menéndez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., and Bunnell, H. T. (1996). *The Nemours Database of Dysarthric Speech*. October 3-6, Proceedings of the Fourth International Conference on Spoken Language Processing, Philadelphia, PA, USA.
- Michie, D., Spiegelhalter, D. J., Taylor, C. C., and Campell, J. (1994). *Machine Learning, Neural, and Statistical Classification*. New York, Ellis Horwood.
- Middag, C., J.-P. Martens, van Nuffelen, G., and De Bodt, M. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing* 2009 2009.
- van der Molen, L., van Rossum, M. A., Ackerstaff, A. H., Smeele, L. E., Rasch, C. R., Hilgers, F. J. (2009). Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views. *BMC Ear, Nose and Throat Disorders* 9(1).
- Nedeljkovic (2002). Image Classification Based On Fuzzy Logic. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 34.
- Nguyen, M. H. (2012). *Segment-based SVMs for Time Series Analysis*. Ph.D. Thesis, Pittsburgh, Pennsylvania, Carnegie Mellon University.

- Nicolosi, L., E. Harryman, Kresheck, J. (2004). *Terminology of Communication Disorders: Speech-Language-Hearing*. Philadelphia, PA: Lippincott, Williams & Wilkins.
- Niedzielska, G. (2001). *Acoustic analysis in the diagnosis of voice disorders in children*. *International Journal of Pediatric Otorhinolaryngology* 57(3): 189–193.
- Nolan, F. (2002). The 'telephone effect' on formants: a response. *Forensic Linguistics* 9(1): 74-82.
- Ogawa, K., H. Yoshihashi, Suzuki, Y., Kamei, S., and Mizutani, T. (2010). Clinical Study of the Responsible Lesion for Dysarthria in the Cerebellum. *International Medicine*, 49(9): 861-864.
- Parncutt, R. and G. E. McPherson (2002). *The science and psychology of music performance*. New York, Schuam Publications.
- Patel, R. (2002). Phonatory control in adults with cerebral palsy and severe dysarthria. *Augmentative Alternative Communication* 18(1): 2-10.
- Pawley, A. (1966). Polynesian languages: A subgrouping based on shared innovations in morphology. *Journal of the Polynesian Society* 75: 39-64.
- Perner, P. and M. Petrou (2003). *Machine Learning and Data Mining in Pattern Recognition*. Springer Verlag.
- Piciarelli, C., C. Micheloni, and Foresti, G. L. (2008). Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and System for Video Technology* 18(11): 1544–1554.
- Platt, L. J., Andrews, G., Young, M., and Quinn, P. T. (1980). Dysarthria of Adult Cerebral Palsy I. Intelligibility and Articulatory Impairment. *Journal of Speech, Language, and Hearing Research* (23): 28-40.
- Potamianos, A. and S. Narayanan (2003). Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing* 11: 603–616.
- Qiao, F., J. Sherwani, Rosenfeld, R. (2010). *Small-vocabulary speech recognition for resource-scarce languages*. First ACM Symposium on Computing for Development.
- Rabiner, L. R. and B. Juang (1993). *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey, Prentice Hall.
- Raschka, S. (2014) *Linear Discriminant Analysis*. Retrieved from http://sebastianraschka.com/Articles/2014_python_lda.html
- Ribarić, S., N. Pavešić, and Zadrija, V. (2009). *Knowledge-Based and Intelligent Information and Engineering Systems*. Lecture Notes in Computer Science Intersection Search for a Fuzzy Petri Net-Based Knowledge Representation Scheme 5711, pp 1-10

- Roberts, R. G., A. Christoffersson, and Cassidy, F. (1989). Real-time event detection, phase identification and source location estimation using single station three-component seismic data. *Geophysical Journal International* 97: 471–480.
- Rosell, M. (2006). An Introduction to Front-End Processing and Acoustic Features for Automatic Speech Recognition.
- Rosen, K. M., Kent, R.D. and Duffy, J. R. (2003). Lognormal distribution of pause length in ataxic dysarthria. *Clinical Linguistics and Phonetics*, 17: 469-486.
- Rosenberg, A. (2009). *Automatic Detection and Classification of Prosodic Events*. Ph.D Thesis, Columbia University.
- Rudzicz, F. (2011). *Acoustic transformations to improve the intelligibility of dysarthric speech*. In Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies, Edinburgh, Scotland.
- Russell, M., D'Arcy, S. and Qun, L. (2007). The effects of bandwidth reduction on human and computer recognition of children's speech. *IEEE Signal Processing Letters*, 14(12): 1044-1046.
- D'Arcy, S. and Russell, M. (2008). *Experiments with the ABI (Accents of the British Isles) Speech Corpus*. Ninth Annual Conference of the International Speech Communication Association
- Sawhney, N. and Wheeler. S. (1999). Using phonological context for improved recognition of dysarthric speech. Technical Report 6345. M. M. Lab.
- Saz, O. (2009). *On-line Personalization and Adaptation to Disorders and Variations of Speech on Automatic Speech Recognition Systems*, Ph.D thesis. Universidad de Zaragoza.
- Saz, O., Simón, J., Rodríguez, W. R., Lleida, E. and Vaquero, C. (2009). Analysis of acoustic features in speakers with cognitive disorders and speech impairments. *EURASIP Journal on Advances in Signal Processing*.
- Saz, O., S.-C. Yin, et al., Lleida, E., Rose, R., Vaquero, C. and Rodríguez, W. R. (2009). Tools and Technologies for Computer-Aided Speech and Language Therapy. *Speech Communication* 51(10): 948-967.
- Schuster, M., Haderlein, T., Nöth, E., Lohscheller, J., Eysholdt, U., Rosanowski, F. (2006). Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *Eur Arch Otorhinolaryngol* 263(2): 188-193.
- Shahin, M., Ahmeda, B., Parnandib, A., Karappab, V., McKechniec, J., Ballardc, K. J., Gutierrez-Osunab, R. (2015). Tabby Talks: An automated tool for the assessment of childhood apraxia of speech. *Speech Communication* 70: 49–64.
- Sharma, H. V. (2008). *Universal Access: Experiments in Automatic Recognition of Dysarthric Speech*, Ph.D thesis. University of Illinois at Urbana-Champaign.

- Shen, V. R. L., Lai, H.-Y. and Lai, A-F. (2013). *Application of High-Level Fuzzy Petri Nets to fall detection system using smartphone*. International Conference on Machine Learning and Cybernetics (ICMLC). Tianjin: 1429 - 1435
- Shriberg, L. D. and Kwiatkowski, J. (1982a). Phonological Disorders II: A conceptual framework for management. *Journal of speech and Hearing Disorders* 47: 242-256.
- Shriberg, L. D. and Kwiatkowski, J. (1982b). Phonological Disorders III: A Procedure for Assessing Severity of Involvement. *Journal of speech and Hearing Disorders* 47: 256-270.
- Sminchisescu, C., Kanaujia, A. and Metaxas, D. (2005). *Conditional models for contextual human motion recognition*. Tenth IEEE International Conference on Computer Vision, 2005, Beijing.
- Strik, H. (2005). The Integration of Phonetic Knowledge in Speech Technology. Text, Speech and Language Technology 25.
- Svadova, M. and Z. Hanzalek (2004). *PN Matlab Toolbox 2.0*, In proceedings of MATLAB conference 2004. Czech Technical University in Prague.
- Sztahó, D., Kis, G., Czap, L., and Vicsi, K. (2014). *A Computer-Assisted Prosody Pronunciation Teaching System*. Proceedings of the 4th Workshop on Child Computer Interaction (WOCCI), Singapore.
- Szwed, P. and M. Komorkiewicz (2013). *Object Tracking and Video Event Recognition with Fuzzy Semantic Petri Nets*. IEEE. Computer Science and Information Systems (FedCSIS), 2013 Federated Conference, Krakow, Poland.
- Taj, S. M. and A. Kumaravel (2015). Survey on Fuzzy Petri Nets for Classification. *Indian Journal of Science and Technology*, 8(14): 1-8.
- Tan, T.-P., Goh, S.-S. and Khaw, Y. M. (2012). *A Malay Dialect Translation and Synthesis System: Proposal and Preliminary System*. International Conference on Asian Language Processing (IALP). Hanoi, Vietnam.
- Thuy, N. T. T., Vien, N. A., Viet, N. H., and Chung, T. C. (2009). Probabilistic Ranking Support Vector Machine. *Advances in Neural Networks, Lecture Notes in Computer Science*, 5552: 345-353
- Ting, H. N., Chia, S. Y., Manap, H. H., Ho, A. H., Tiu, K. Y., Abdul Hamid, B. (2012). Fundamental Frequency and Perturbation Measures of Sustained Vowels in Malaysian Malay Children Between 7 and 12 Years Old. *Journal of Voice* 26(4): 425-430.
- Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York, SpringerVerlag.

- Vipperla, R., Renals, S., and Frankel, J. (2010). Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing* 2010(5).
- Wertzner, H. F., Schreiber, S. and Amaro, L. (2005). Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders. *Brazilian Journal of Otorhinolaryngology* 71(5): 582-588.
- White, K. (2012). *Acoustic characteristics of Ataxic Dysarthria*. Honors. The Department of Speech, Language and Hearing Sciences, University of Florida.
- White, R. L. (2000). Object Classification as a Data Analysis Tool. *Astronomical Data Analysis Software and Systems*. 216.
- Whitehill, T. L. and V. Ciocca (2000). Perceptual-phonetic predictors of single word intelligibility: A study of Cantonese dysarthria. *Journal of Speech Language and Hearing Research*, 43(6)
- Wilcox, K. A. and Y. Horii (1980). Age and changes in vocal jitter. *Journal of Gerontology*. 35(2): 194–198.
- Wilson, C. Y. E. (2004). *Articulatory strategies, speech acoustics and variability*. From sound to sense: 50+ years of discoveries in speech communication, MIT, Cambridge.
- Yamato, J., J. Ohya, and Ishii, K. (1992). Recognizing human action in time sequential images using hidden Markov model. *Computer Vision and Pattern Recognition*.
- Yorkston, K. M., Beukelman, D. R. and Tice (1996). Sentence Intelligibility Test. Lincoln, NE, Tice Technology Services, Inc.
- Young, S., G. Evermann, Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. (2006). HTK Book, Cambridge University Engineering Department.
- Young, V. and Mihailidis, A. (2010). Difficulties in Automatic Speech Recognition of Dysarthric Speakers and Implications for Speech-Based Applications Used by the Elderly: A Literature Review. *Assistive Technology*, 22(2): 99–112.