

**UNIVERSITI TEKNOLOGI MARA**

**AUTOMATED SENTENCE  
BOUNDARY DETECTION FOR  
SPONTANEOUS SPEECH IN MALAY  
LANGUAGE**

**MUHAMMAD IZZAD BIN RAMLI**

Thesis submitted in fulfillment  
of the requirements for the degree of  
**Master of Science**

**Faculty of Computer and Mathematical Sciences**

**November 2013**

## AUTHOR'S DECLARATION

I declare that the work in this thesis entitled "Sentence Boundary Detection for Spontaneous Speech in Malay Language" was carried out in accordance with the regulation of Universiti Teknologi MARA. It is original and the result of my own research except as cited in the references. This thesis has not been submitted to any other academic institution for any other degree of qualification.

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

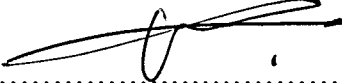
Name of Student : Muhammad Izzad bin Ramli

Student I.D.No : 2011289094

Programme : Master of Sciences (CS 780)

Faculty : Computer and Mathematical Sciences

Thesis Title : Automated Sentence Boundary Detection for Spontaneous Speech in Malay Language

Signature of Student :  .....

Date : November 2013

## ABSTRACT

Sentence boundary detection (*SBD*) or also known as sentence breaking decides where sentences begin and end. Sentence boundary detection is necessary in many applications, such as speech summarization, video summarization, speech document indexing and retrieval. This research describes sentence boundary detection in spontaneous Malay language spoken audio. Spontaneous speech is a speech that is not planned or arranged beforehand. Related speech studies for spontaneous Malay language speech are still lacking and no work has been done on sentence boundary. Previous studies showed that combination of linguistic and acoustic approach for sentence boundary detection is able to provide better than using only one approach. However, linguistic model for Malay language is still not available, only acoustic approach is used for Malay language sentence boundary detection. Therefore, the combination of prosodic features with volume features and rate-of-speech (*ROS*) was proposed for sentence boundary detection of spontaneous speeches. The data used are from spontaneous speeches of Malaysian Parliament Hansard Document (*MPHD*). Experiments are conducted on 42 minutes of Malay language spontaneous speeches comprising of 6,413 speech and non-speech segments. Then, non-speech segments are selected as the candidates for the sentence boundary detection experimental data. The accuracy achieved for the proposed speech and non-speech detection method is 97.8% and the sentence boundary detection is 100% with false alert 19.44%. As the outcome, the proposed methods of sentence boundary detection using fusion of prosodic features, volume and rate-of-speech (*ROS*) and Adaboost managed to detect and label sentence boundary automatically.

## ACKNOWLEDGEMENTS

Alhamdulillah and I am grateful to Allah S.W.T for His blessing and mercy that enable me to complete my research dissertation on time.

During the project progresses, many people had help and give support. I wish to thank and express my biggest appreciations to them especially to my supervisor, Assoc. Prof. Dr. Nursuriati Jamil for her determination in giving me continuous guidance, precious advices throughout the duration of the research. Special thanks to my second supervisor, Prof. Dr Zainab Abu Bakar for her assistance, guidance and support for my research dissertation.

Not to forget to all my lecturers and my friends who have been very supportive in sharing ideas and aiding me in completing this research dissertation. Lastly, thanks to my beloved parents and my supportive family who have never stop giving me full support, understanding and courage throughout the research dissertation. Not to forget everyone who are involve directly or indirectly in the completion of this proposal.

Thank You.

# TABLE OF CONTENTS

	<b>Page</b>
<b>AUTHOR'S DECLARATION</b>	ii
<b>ABSTRACT</b>	iii
<b>ACKNOWLEDGEMENTS</b>	iv
<b>TABLE OF CONTENTS</b>	v
<b>LIST OF TABLES</b>	ix
<b>LIST OF FIGURES</b>	xi
<b>CHAPTER ONE: INTRODUCTION</b>	
1.1 Background	1
1.2 Problem Statements	2
1.3 Objectives	3
1.4 Contributions of the Research	4
1.5 Scope of Project	5
1.6 Significant of Study	5
1.7 Organization of the Thesis	6
1.8 Conclusion	7
<b>CHAPTER TWO: LITERATURE REVIEW</b>	
2.1 Introduction	8
2.2 Sentence Boundary Detection	8
2.3 Sentence Boundary Techniques	14
2.3.1 Acoustic Approach	14
2.3.1.1 Pitch Break Behaviour	17
2.3.1.2 Pause Duration Model	18
2.3.1.3 The Sorted Pitch Break Map	18