Technical University of Liberec

Faculty of Mechatronics, Informatics and Interdisciplinary Studies

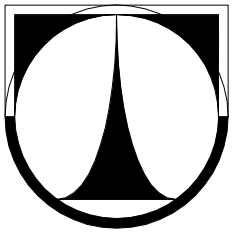
Doctoral Thesis

June 2014 **Nguyen ThienChuong**

Technical University of Liberec

Faculty of Mechatronics, Informatics and Interdisciplinary Studies

Institute of Information Technology and Electronics



Automatic speech recognition of Vietnamese

Nguyen Thien Chuong

Doctoral Thesis

Supervisor: doc. Ing. Josef Chaloupka, Ph.D.

Ph.D. Program: P 2612 Electronics and Informatics

Branch of Study: 2612V045 Technical Cybernetics

Declaration

I declare herewith that this Ph.D. thesis is my own work and that all used resources are listed in the Bibliography section.

Nguyen Thien Chuong

Liberec, June 6, 2014

Acknowledgements

Herewith I would like to thank my supervisors, doc. Ing. Josef Chaloupka, Ph.D. firstly, for an interesting and challenging thesis topic, and secondly, for the guidance, advices and patience he provided.

I would also like to give special thanks to prof. Ing. Jan Nouza, CSc. for his help and support during the time of my Ph.D. studies.

I am very grateful to all colleagues for their support and kindness. Thanks to Karel Paleček and Jiří Málek for the friendship and cooperation in the time of my studies.

The thesis could not be completed without the support of the Department of Information Technology and Electronics (ITE) at the Faculty of Mechatronics, Informatics and Interdisciplinary Studies. Financial support for the work was provided through the Student Grant Scheme (SGS) at the Technical University of Liberec.

Abstract

This thesis presents the work on automatic speech recognition of Vietnamese, a tonal, syllable-based language in which new approaches have to be applied to obtain reliable results. When dealing with Vietnamese, the following basic problems have to be solved or clarified: the selection of phonetic unit to build acoustic models, the collection of text and speech corpora, the creation of pronouncing dictionary, the construction of language model and especially, the methods to deal with tone.

With the basic idea of systematically and methodically finding solutions to all the problems mention above, in this work, several methods for collecting large text and speech corpora are first described in which two types of text corpora are obtained by exploiting the source of linguistic data from the Internet, and also two types of speech corpora are extracted, including an Internet-based large continuous speech corpus and a recorded audio-visual speech corpus. Then, a standard phoneme set optimal to Vietnamese with its corresponding grapheme-to-phoneme mapping table is proposed. By constructing various types of pronunciation dictionaries and language models for Vietnamese, the optimal way to integrate tone in a syllable as well as the strategies to deal with speech recognition of Vietnamese will be totally examined in the form of large vocabulary continuous speech recognition tasks.

The study is further extended to the field of audio-visual speech recognition of Vietnamese in which the performance gains of audio only speech recognition in noisy condition is proved to be noticeable when integrated with visual information. In this work, many types of visual front ends and visual features are examined in the task of isolated-word speech recognition of Vietnamese.

Contents

Acknowledgements ..... **vii**

Abstract ..... **ix**

Contents ..... **xi**

List of Figures ..... **xiii**

List of Tables ..... **xv**

Glossary ..... **xix**

Chapter 1 Introduction **1**

1.1 Text and speech corpora 2
1.2 Basic problems of LVCSR of Vietnamese 2
1.3 Audio-visual speech recognition 3

Chapter 2 Automatic speech recognition **5**

2.1 The basis of automatic speech recognition 5
2.2 History of ASR technology 11

Chapter 3 Vietnamese Language and Speech Recognition Studies **19**

3.1 Introduction of Vietnamese 19
3.2 Vietnamese phonology 21
3.2.1 Vowels 21
3.2.2 Consonants 23
3.2.3 Tones 24
3.2.4 Syllable 26
3.3 State of the art ASR of Vietnamese 29

Chapter 4 Thesis Goals **47**

Chapter 5 Strategies for Speech Recognition of Vietnamese **49**

5.1 Phoneme set proposal 49
5.2 Pronunciation dictionary creation 51
5.3 Strategies for speech recognition of Vietnamese 54
5.3.1 Phoneme-based strategy 54
5.3.2 Vowel-based strategy 58
5.3.3 Rhyme-based strategy 62
5.3.4 Syllable-based strategy 64

Chapter 6 Text, Audio and Audio-Visual Databases **67**

6.1 Building of Vietnamese text corpus from the Internet 67
6.1.1 Vietnamese text corpus from Wikipedia 67

6.1.2 Extracting of general purpose text corpus 68
6.1.3 Extracting of specific purpose text corpus 70
6.1.4 Filtering of text corpora 71
6.2 Collecting of speech corpus for LVCSR task 73
6.3 Designing of audio-visual speech corpus 73

Chapter 7 Audio Speech Recognition of Vietnamese 77

7.1 Building language model for LVCSR of Vietnamese 77
7.1.1 Syllable-based LM construction 78
7.1.2 Word-based LM construction 82
7.2 Isolated word speech recognition 84
7.3 Experiments on LVCSR of Vietnamese 86
7.3.1 Examining the Effect of Tone in Vietnamese Syllables 86
7.3.2 Context-dependent LVCSR of Vietnamese 91
7.3.3 The effect of LM on LVCSR of Vietnamese 92
7.3.4 Gender-dependent LVCSR of Vietnamese 95

Chapter 8 Audio-Visual Speech Recognition of Vietnamese 97

8.1 Introduction 97
8.2 Feature extraction 98
8.2.1 Face and facial features localization 98
8.2.2 Region of interest extraction 100
8.2.3 LDA data projection 101
8.2.4 Visual front end for feature extraction 103
8.3 Isolated word visual only speech recognition 104
8.4 Isolated word audio-visual speech recognition. 107
8.4.1 Audio-visual integration 107
8.4.2 Audio-visual fusion experiments 110

Chapter 9 Conclusion 123

9.1 Thesis achievements 123
9.2 Text and speech corpora 123
9.3 Tone hypotheses 124
9.4 Audio speech recognition of Vietnamese 125
9.5 Audio-visual speech recognition 127
9.6 Future work 128

Appendix A Basic Phonetic Unit Set 129

A.1. Phoneme set of phoneme-based strategy 129
A.2. Phoneme set of vowel-based strategy 132
A.3. Phoneme set of rhyme-based strategy 138

Appendix B Audio visual experiments 147

B.1. Weight selection for white noise 147
B.2. Weight selection for babble noise 149
B.3. Weight selection for Volvo noise 151

Bibliography 153

List of Figures

Fig.2.1: Components of an ASR system.....10
Fig.2.2: Feature vector extraction from speech signal.....11
Fig.3.1: The Vietnamese alphabet and tone in writing system.....20
Fig.3.2: Pitch contours and duration of the six Northern Vietnamese tones as uttered by a male speaker (not from Hanoi). Fundamental frequency is plotted over time [96].....25
Fig.3.3: Vietnamese syllable's structure.....26
Fig.5.1: Analysis of syllable 'TOÁN' in phoneme-based strategy. 55
Fig.5.2: Analysis of syllable 'TOÁN' in vowel-based strategy. 59
Fig.5.3: Analysis of syllable 'TOÁN' in rhyme-based strategy.....62
Fig.5.4: Analysis of syllable 'TOÁN' in syllable-based strategy.....65
Fig.6.1: Block diagrams of sentence selection procedures: a) Sentence selection block 1; b) Sentence selection block 2. 75
Fig.7.1: Constructing and testing *n*-gram LM. 77
Fig.7.2: The effect of number of states on various feature types. 85
Fig.8.1: The Architecture of CLM.....99
Fig.8.2: ROI extraction.....100
Fig. 8.3 : Across frame features selection. 102
Fig. 8.4: Assigning of feature vectors to the audio classes.....103
Fig.8.5: Visual front end feature extraction.....104
Fig.8.6: HLDA and 1-Stage LDA for DCT coefficient with WS = 7. 106
Fig.8.7: HLDA for DCT, PCA and AAM using the best WS for each feature type. .107
Fig.8.8: Recognition results for audio only (AO) and visual only (VO) using additive noises. 112
Fig.8.9: Recognition results for audio only (AO), visual only (VO) and audio-visual (MI) in white noise condition. 112
Fig.8.10: Recognition results for audio only (AO), visual only (VO) and audio-visual (MI) in babble noise condition.....113

Fig.8.11: Recognition results for audio only (AO), visual only (VO) and audio-visual (MI) in volvo noise condition..... 113

Fig.8.12: MI using adaptation data (WA) compare to equal weight (W11) in white noise condition. 114

Fig.8.13: LI using exhausted search strategy (LI WA) in white noise condition. 116

Fig.8.14: LI using exhausted search strategy (LI WA) in babble noise condition. 116

Fig.8.15: LI using exhausted search strategy (LI WA) volvo noise condition. 117

Fig.8.16: LI using confidence score strategy in white noise..... 119

Fig.8.17: LI using confidence score strategy in babble noise..... 119

Fig.8.18: LI using confidence score strategy in volve noise condition..... 120

Fig.8.19: Comparison of fusion strategies in white noise..... 120

Fig.8.20: Comparison of fusion strategies in babble noise. 121

Fig.8.21: Comparison of fusion strategies in volvo noise..... 121

Fig.B.1: LI using N-best dispersion score with different weights *w* in white noise. .. 147

Fig.B.2: LI using Variance score with different weights *w* in white noise..... 148

Fig.B.3: LI using N-best average score with different weights *w* in white noise. 148

Fig.B.4: LI using N-best dispersion score with different weights *w* in babble noise. 149

Fig.B.5: LI using Variance score with different weights *w* in babble noise. 150

Fig.B.6: LI using N-best average score with different weights *w* in babble noise. 150

Fig.B.7: LI using N-best dispersion score with different weights *w* in volvo noise.. 151

Fig.B.8: LI using Variance score with different weights *w* in volvo noise..... 152

Fig.B.9: LI using N-best average score with different weights *w* in volvo noise..... 152

List of Tables

Tab. 3.1: Examples of Vietnamese word and their form. 20
Tab. 3.2: IPA chart of monophthongs. 21
Tab. 3.3: Pronunciations of 12 vowel letters. 22
Tab. 3.4: Combination of two vowels in Vietnamese. 22
Tab. 3.5: Combination of three vowels in Vietnamese. 23
Tab. 3.6: IPA chart of Vietnamese consonants. 23
Tab. 3.7: Pronunciations of consonant letters. 24
Tab. 3.8: Six Vietnamese tones. 25
Tab. 3.9: Syllable's type. 26
Tab. 3.10 Text corpus size and perplexity of the language models. 30
Tab. 3.11: The size of corpus collected from Vietnamese electronic documents. 31
Tab. 3.12: Vietnamese GlobalPhone speech corpus 34
Tab. 3.13: The VOV speech corpus. 35
Tab. 3.14: Evaluation of the VOV language models. 36
Tab. 3.15: Evaluation of the language models. 36
Tab. 3.16 : Language models perplexities. 37
Tab. 3.17: Text corpus for construction of statistical LM. 37
Tab. 3.18: Perplexity of Vietnamese LMs on the test corpus 37
Tab. 3.19 : Syllable-based and word-based perplexities. 38
Tab. 5.1: 39 English Phonemes in CMU pronunciation dictionary. 49
Tab. 5.2: Grapheme-to-phoneme mapping table. 50
Tab. 5.3: Vietnamese syllable analyzing schemes. 52
Tab. 5.4: Vietnamese pronunciation dictionary types. 53
Tab. 5.5: Number of possible basic phonetic unit. 53
Tab. 5.6: Analyzing of syllable 'TOAN' using C1wVC2 method. 56
Tab. 5.7: Analyzing of syllable 'TOAN' using C1wVC2T_I, C1wVTC2_I and C1wVTC2T_I methods 57

Tab. 5.8: Analyzing of syllable ‘TOAN’ using C1wVC2T_D, C1wVTC2_D and C1wVTC2T_D methods. 58
Tab. 5.9: Analyzing of syllable ‘TOAN’ using C1MC method..... 60
Tab. 5.10: Analyzing of syllable ‘TOAN’ using C1MCT_I and C1MTC_I methods.. 61
Tab. 5.11: Analyze syllable ‘TOAN’ using C1MCT_D and C1MTC_D methods. 61
Tab. 5.12: Analyzing of syllable ‘TOAN’ using C1R method. 63
Tab. 5.13: Analyze syllable ‘TOAN’ using C1RT_I method. 64
Tab. 5.14: Analyzing of syllable ‘TOAN’ using C1RT_D method. 64
Tab. 6.1: Statistics of Wikipedia text corpus. 68
Tab. 6.2: Example of seed words. 69
Tab. 6.3: Statistics of query length selection. 70
Tab. 6.4: Statistics of raw data. 70
Tab. 6.5: Website for collecting text corpus. 71
Tab. 6.6: Statistics of the filtered general purpose text corpus. 72
Tab. 6.7: Statistics of the filtered specific text corpus (news). 72
Tab. 6.8: Statistics of the filtered specific text corpus (literature). 72
Tab. 6.9: Statistics of the total filtered text corpora. 72
Tab. 6.10: Original sentence set statistics. 74
Tab. 6.11 : 50 isolated words for audio-visual speech data recording. 76
Tab. 7.1: LM test on general purpose text corpus. 79
Tab. 7.2: LM test on specific text corpus (literature). 79
Tab. 7.3: LM test on specific text corpus (news). 79
Tab. 7.4: LM test on total text corpus. 80
Tab. 7.5: LM test using Good-Turning smoothing. 81
Tab. 7.6: LM test using Witten-Bell smoothing. 81
Tab. 7.7: LM test using Kneser-Ney smoothing. 81
Tab. 7.8: LM test using Kneser-Ney smoothing with interpolation. 81
Tab. 7.9: LM test using vocabulary of 6000 syllables. 81
Tab. 7.10: LM test using vocabulary of 7000 syllables. 82
Tab. 7.11: LM test using vocabulary of all syllables (11017). 82
Tab. 7.12: LM test using vocabulary of 5741. 82

Tab. 7.13: Multi-syllable-based bi-gram LM test.....84
Tab. 7.14: Recognition rate [%] for isolated word speech recognition. 86
Tab. 7.15: Speech corpus for LVCSR tasks. 86
Tab. 7.16: SACC [%] for context-independent LVCSR. 90
Tab. 7.17: SACC [%] for context-dependent LVCSR. 91
Tab. 7.18: SACC [%] for various text corpus categories. 92
Tab. 7.19: Percent of text covered by syllables in vocabulary. 93
Tab. 7.20: SACC [%] for various smoothing method and vocabulary size. 93
Tab. 7.21: SACC [%] for multi-syllable-based LMs.....94
Tab. 7.22: SACC [%] for gender-dependent recognizers.....96
Tab. 8.1: Analyzing of Vietnamese syllabel into basic units. 102
Tab. 8.2: Recognition results (VI) for various visual parameters using inner frame LDA. 105
Tab. 8.3: Recognition results for various methods of Vietnamese syllable analysis using inner frame LDA (VI). 106
Tab. 8.4: Recognition results using HLDA (VH) and 1-Stage LDA (VS) with different WS. 106
Tab. 8.5: Recognition Results Using HLDA with Different Types of Visual Feature. 107
Tab. 8.6: recognition results for MI using both stream weights = 1 and using the best stream weights for each SSNR. 115
Tab. 8.7: N-best hypotheses for each confidence score type.....118

Glossary

ACC	word accuracy
AAM	active appearance model
ASM	active shape model
ASR	automatic speech recognition
AVSR	audio-visual speech recognition
CLM	constrained local models
CMN	cepstral mean normalization
CSR	continuous speech recognition
DCT	discrete cosine transform
DFT	Discrete Fourier Transform
DWT	discrete wavelet transform
Fps	frames per second
GMM	Gaussian mixture model
HLDA	hierarchical linear discriminant analysis
HMM	hidden Markov model
HTK	a hidden Markov model toolkit
IPA	International Phonetics Association
LDA	linear discriminant analysis
LM	language model
LPC	linear prediction filter coefficient
LVCSR	large vocabulary continuous speech recognition
MFCC	Mel frequency cepstral coefficient
MLLT	maximum likelihood linear transform
PCA	principal component analysis
PDM	point distribution model
ROI	region of interest
SACC	syllable accuracy

SER	syllable error rate
SNR	signal to noise ratio
SSNR	segmental signal to noise ratio
SVM	support vector machine
WER	word error rate

*** Experiments and Evaluation**

C1wVC2	phoneme-based scheme, no tone
C1wVC2T_D	phoneme-based scheme, dependent tone at the end of syllable
C1wVC2T_I	phoneme-based scheme, independent tone at the end of syllable
C1wVTC2_D	phoneme-based scheme, dependent tone after main vowel
C1wVTC2_I	phoneme-based scheme, independent tone after main vowel
C1wVTC2T_D	phoneme-based scheme, two dependent tones located after main vowel and at the end of syllable
C1wVTC2T_I	phoneme-based scheme, two independent tones located after main vowel and at the end of syllable
C1MC	vowel-based scheme, no tone
C1MCT_D	vowel-based scheme, dependent tone at the end of syllable
C1MCT_I	vowel-based scheme, independent tone at the end of syllable
C1MTC_D	vowel-based scheme, dependent tone after vowel
C1MTC_I	vowel-based scheme, independent tone after vowel
C1R	rhyme-based scheme, no tone
C1RT_D	rhyme-based scheme, dependent tone at the end of syllable
C1RT_I	rhyme-based scheme, independent tone at the end of syllable
S	syllable-based scheme, no tone
ST_I	syllable-based scheme, independent tone at the end of syllable
ST_D	syllable-based scheme, dependent tone on the whole syllable

Chapter 1

Introduction

In recent years, with the development of computer technology, many works related to human speech were made to be applicable in practice. Some of the prominent areas which based on the above works consist of speech synthesis, speech compression, speech recognition, speaker identification, etc. In these areas, speech recognition has extracted interest from many researchers in which noticeable successes were obtained, especially in the designing of large vocabulary continuous speech recognition (LVCSR) systems. The final aim of every speech recognition research is the construction of a system which enables the natural communication by speech from people to machines. Such system is needed because speech is human most natural mode of communication. In addition, speech provides the highest potential speed in human-to-machine communication since people speak much faster than when they write or type. Also, speech recognition systems free the eyes and hands of the operator to perform other tasks simultaneously.

The research on automatic speech recognition (ASR) of Vietnamese has made significant progress since it was first introduced more than twenty years ago. However, ASR of Vietnamese is just at its experimental stage and yet to reach the performance level required to be widely used in real-life applications. Incoherence is one of the prominent characteristic of works related to ASR of Vietnamese in which there is still not any standard database or method to deal with speech recognition tasks. Researchers in this field usually propose their own database and method to solve a given ASR problem, and so this makes the cooperation among researchers become very difficult or impossible.

Motivated by the successes of modern speech recognition systems as well as the development of ASR of Vietnamese, an under-resourced language, this work is dedicated to provide the basic ideas, hypotheses and methods for dealing with Vietnamese language, which can be used as baseline methodology for all the future works on ASR of Vietnamese.

1.1 Text and speech corpora

An important problem when constructing ASR systems is the presence of text and speech corpora of suitable size. These corpora are not available for ASR of Vietnamese, especially for LVCSR or audio-visual speech recognition tasks.

In this work, a significant amount of time is used to collect two types of data. The first data is a collection of text and speech corpora from the Internet resource for LVCSR task in which the speech corpus is manually segmented and transcribed to obtain a reasonable large number of good utterances. The second data is an audio-visual speech corpus which is recorded in controlled room condition. This corpus contains both isolated word and continuous speech that is used to evaluate audio-visual isolated word speech recognition task.

1.2 Basic problems of LVCSR of Vietnamese

For Vietnamese, there are several obstacles one has to deal with when constructing a speech recognition system. This thesis is mainly concerned with the following basic problems:

The first problem is the proposal of a phoneme set optimal to Vietnamese. It is noteworthy that a standard phoneme set for Vietnamese is not available. In many works, graphemes are used in place of phonemes which are straightforward to obtain. In some other works, a phone set is presented but a standard phoneme set is not proposed. In this work, both grapheme-based and phone-based phoneme set are proposed and evaluated in the form of LVCSR of Vietnamese.

The second problem is the construction of a pronunciation dictionary. But again, a standard pronunciation dictionary is not available for works related to ASR of Vietnamese. This thesis considers four main strategies including phoneme-based, vowel-based, rhyme-based and syllable-based to deal with this problem. Each strategy has different set of phonetic units and is compared to other strategies on the same speech recognition task.

The final, and also the most interesting problem when dealing with ASR of Vietnamese is the interpretation of hypotheses about tone. Is tone a dependent component? Where is the position of tone in a syllable? What is the effect of different tone’s

hypotheses? All of these questions will be clarified in the task of context-dependent and context-independent LVCSR of Vietnamese.

1.3 Audio-visual speech recognition

With the aim of building a command control system, this thesis is also concerned with audio-visual speech recognition in the form of an isolated word task. First, to select the best visual front-end for feature extraction, two different visual front-ends are considered in which various visual feature types are evaluated and compared. Using the best feature and visual front-end, the final evaluation is then performed by integrating the auditory and visual streams into the final recognition system. Two fusion strategies are examined for the most successful visual feature type selected.

Chapter 2

Automatic speech recognition

2.1 The basis of automatic speech recognition

Automatic speech recognition can be defined as a process of transcribing a spoken speech signal of a specific language into a sequence of words in readable text format by using algorithms implemented as a computer program.

Making a computer understand and respond properly to fluently spoken speech has attracted researchers for more than six decades. Many important progressions in ASR technology have been obtained for the last several years in which ASR systems with vocabulary sizes exceeding 65000 words using fast decoding algorithms allow continuous speech recognition process approaching near real-time response [1]. Although ASR technology is becoming more and more popular in a number of applications and services such as voice dictation, home automation, automatic information access (travel, banking, etc.), automatic processing in telephone networks, etc., it is not yet at the level where computers can understand every spoken word, in any speaking environment, or by any speakers.

So, what makes ASR so difficult? When communicating to each other, humans use not only information hearing from their ears but also other signals from speaker's body such as facial expressions, postures, hand gestures, etc. They also use the knowledge or information about the speakers and the subjects, which is totally missed by ASR systems. Many attempts have been done to model this knowledge but the question is how much it is needed in an ASR system to obtain human comprehension? Uttered speech always contains unwanted information called noise. Noise can be sound of any kind, a car running, a computer fan humming, a clock ticking, a song background, etc. Identifying, tracking and filtering out these noises from the speech signal are also a big challenge. Another difficulty is the variability of channel. Here one faces the problem of speech distortion, echo effect (a

phenomenon where a speech signal bounced on some object and come back to the microphone), various type of transducer (microphone, telephone, etc.) and other effects that change the discrete representation of a speech signal in a computer. What is about speaker's variability? Every speaker is a unique individual with his or her own physical body and personality. The voice uttered by a speaker can be from man or woman; from kid, adult or the elderly; from strong or weak one, etc. A speaker is different from other speakers not only in his or her physical attributes of the body such as the lung size, the size and shape of the vocal cord, the formation of the cavity or palate, etc., but also in his or her region of living or social standing which contribute to their specific speaking styles (personal vocabulary, way to pronounce and emphasize, situations of communication, etc.). As a result, speech uttered by a speaker is special. On the other hand, variations in the voice can also occur within one specific speaker. One virtually cannot pronounce exact the same word even if he or she tries to do it over and over again. Speech produced when you are happy will be different from when you are disappointed, stressed, sad or frustrated. This difference occurs not only in the power containing in speech but also in the speed of speech. Another problem causing difficulty is ambiguity in natural spoken speech. One source of this ambiguity is homophones where words are pronounced the same but have different orthography. The other source is word boundary ambiguity. This occurs in continuous speech in which a set of words is put together into a sentence. So we can see that there are many problems one has to concern when building an ASR system. Can ASR obtain the level of natural human communication? May be not, but progressions in the last several years show that constructing a good enough ASR system is not impossible.

Many researches on ASR have been done in which they cover a large range of applications and can be in general classified as follows:

- Base on the properties of input speech, ASR systems can be classified into isolated word and continuous speech recognition task. In continuous speech recognition, the system has to recognize sequence of words of a given speech signal. This kind of system is complex because of incomplete representation of all possible input patterns, and so they have to use patterns of smaller speech events (phones) to explain larger sequences (sentences, paragraphs, etc.). The isolated word systems, on the other hand, are easier to

construct and must more robust than the continuous speech recognition systems as they have the complete set of patterns for all possible inputs.

- When viewing in the speaker property aspect, these systems are split into speaker dependent and speaker independent speech recognition tasks. In speaker dependent systems [2], the models are trained or adapted for a single speaker, and so, they can only understand speech uttered by that specific speaker. For the systems to understand other speakers, new models have to be trained or adapted using speech data specified for these speakers. Systems of this kind are more feasible for personal purposes, i.e. dictation system used on personal computer, because user is asked to perform an hour or so to complete the training process. For speaker independent systems, they have to handle many speakers, and so, the models are trained just one time for all speakers. Speaker independent systems are not as accurate and stable as speaker dependent systems, but they are more feasible for general purposes, i.e. in automated telephone operator system.

- With respect to the vocabulary size, ASR can be classified into small, medium and large vocabulary speech recognition task. In general, the bigger the vocabulary size is the more complicated the ASR task will be. Tasks with vocabulary size less than 100 words are typically classified as small vocabulary task [3-5]. For this type of tasks, high recognition rate can easily be achieved for a wide range of speakers. Large vocabulary tasks are tasks with more than 20,000 words [6] (for syllable-based language like Vietnamese, ASR tasks with vocabulary of size more than 5000 syllables can be considered as large vocabulary tasks) in which high accuracy can be obtained with speaker dependent property. For medium size vocabulary tasks, the size of the vocabulary is on the order of 1000 to 3000 words.

For applications of the above types, there are three basic approaches to deal with speech recognition tasks: acoustic-phonetic approach, pattern recognition approach and artificial intelligence approach.

- The acoustic-phonetic approach is the earliest one which uses knowledge of phonetics and linguistics to guide the decoding process [7]. The core of this approach is the definition of a set of rules (phonetics, phonology, phonotatics, syntax, pragmatics, etc.) that might help to decode speech signal. There are three main steps for this approach: (1) speech analysis and feature detection, (2) segmentation and labelling, and (3) determining of valid

word (sentence) from the phonetic label sequences. Although having strong theoretical base, works based on this approach usually give poor results because of the lack of a good knowledge of acoustic phonetics and other related areas. These difficulties arise from extracting proper acoustic properties for features, expressing phonetic rules, making rules interact, improving the system, etc.

- The pattern recognition approach [8, 9] contains two basic steps called pattern training and pattern matching. In the training step, a mathematical representation of a specific speech pattern (phone, word, or phase) can be constructed in form of a speech template or a statistical model. In the matching step, an unknown speech signal is compared to all possible learned patterns in the previous step to classify it into a given label corresponding to a specific pattern. The pattern recognition approach is widely used for speech recognition in the last several decades and contains two large branches: template-based approach and stochastic approach.

The basic idea of template-based approach for speech recognition is as follows: given a set of N trained templates $T = [t_1, t_2, \ldots, t_N]$, a concatenated sequence of templates R^S is a subset of S templates taken from T. The recognition is a process of finding the best word sequence W that minimizes a distance function between observation sequence O and a sequence of reference templates. So the problem is to find the optimum sequence of templates R^* that best matches O, as follows,

$$R^* = \arg\min_{R^S} d(R^S, O). \qquad (2.1)$$

For this approach, the complexity will grow exponentially with the length of the sequence of words W in which it becomes computationally expensive or impractical to implement. In addition, the sequence of templates does not take into account the silence or the coarticulation between words. This approach is usually applied on word level because it avoids the segmentation and classification error which can occur on phone level. Dynamic time warping (DTW) [10, 11] and vector quantization (VQ) [12-15] are two widely used methods in speech recognition tasks specified for template-based approach. DTW is a well-known technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions. Intuitively, the sequences are warped in a nonlinear fashion to match each other. This method is used in ASR to cope with the difference in speaking speeds of speech patterns. VQ is one of the most efficient source-coding

techniques in which it encodes the speech patterns from a set of possible words (continuous signals) into a smaller set of vectors (discrete symbols) to perform pattern matching. A vector quantizer is described by a codebook, which is a set of fixed prototype vectors or reproduction vectors (codeword). To perform the quantization process, the input vector is matched against each codeword in the codebook using some distortion measure. The input vector is then replaced by the index of the codeword with the smallest distortion.

For stochastic approach, probabilistic model is used to deal with uncertain or incomplete information. It can be seen as an extension of template-based approach using more powerful mathematical and statistical tools. In this framework, the decoder attempts to find the sequence of words *W* which is most likely to have generated acoustic vectors *O*, i.e. the decoder tries to find

$$\widehat{W} = \arg\max_{W}\{P(W \mid O)\}. \qquad (2.2)$$

However, since $P(W \mid O)$ is difficult to model directly, Bayes' Rule is used to transform (2.2) into the equivalent problem of finding:

$$\widehat{W} = \arg\max_{W}\{P(O \mid W)P(W)\}. \qquad (2.3)$$

The likelihood $P(\mathrm{O} \mid \mathrm{W})$ is determined by an acoustic model and the prior probability $P(W)$ is determined by a language model.

Hidden Markov model (HMM) [16] is the most popular stochastic model which is used in almost every modern speech recognition applications. A HMM is characterized by a finite-state Markov model and a set of output distributions. The reason for the popularity of HMM is the existence of several elegant and efficient algorithms.

- The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. It involves two basic ideas. First, it studies the thought processes of human beings. Second, it deals with representing those processes via machines (computers, robots, etc.). With the development of computer technology at current time, artificial intelligence approach is becoming more and more popular and can gradually compete or even overcome the other approaches.

Because of the power of stochastic approach, almost any modern speech recognition systems are built based on its framework. The basic structure of statistical framework for an ASR system consists of the following main components (Fig.2.1):

- Preprocessing block: the block where speech signal is improved using some preprocessing steps. First, the signal is spectrally flatten using a first order high pass FIR filter to emphasize the higher frequency components. The signal is then divide into frames with an appropriate time length for each frame, and a Hamming window is applied for each frame to reduce the signal discontinuity at the end of each block.

- Feature extraction block: this block is used to extract a set of features from the speech signal which contain the most useful information for the classification task. These features have to be sensitive to linguistic content and robust to acoustic variation, or more specifically, the selected features can be distinguished between different linguistic units (e.g. phones). Also, the features should be robust to noise and other factors that are irrelevant for the recognition process. Fig.2.2 describes the way feature vectors are extracted from a speech signal. In ASR systems, various methods have been used for feature extraction task such as principal component analysis (PCA), linear discriminant analysis (LDA), independent component analysis (ICA), linear predictive coding, cepstral analysis, mel-frequency scale analysis, filter bank analysis, mel-frequency cepstrum, etc.

- Classification block: acoustics models, pronunciation dictionary and language model are the main components of this block. The acoustics models are usually Hidden Markov Models (HMMs) trained for whole words or phones as linguistic units. The pronunciation dictionary defines the appropriate combination of phones for a valid word. And the language model is used to predict the likelihood of specific words occurring in sequence in a certain language.

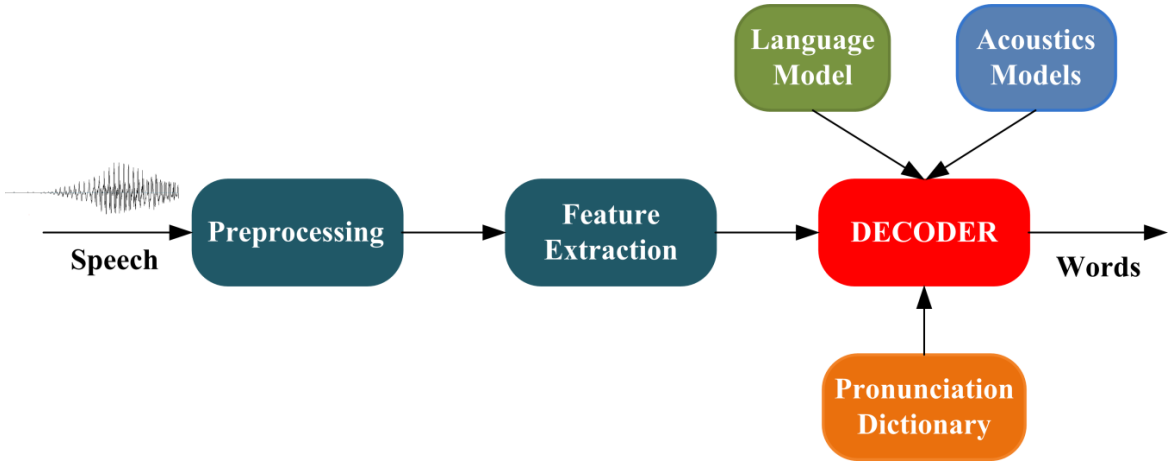


Fig.2.1: Components of an ASR system.

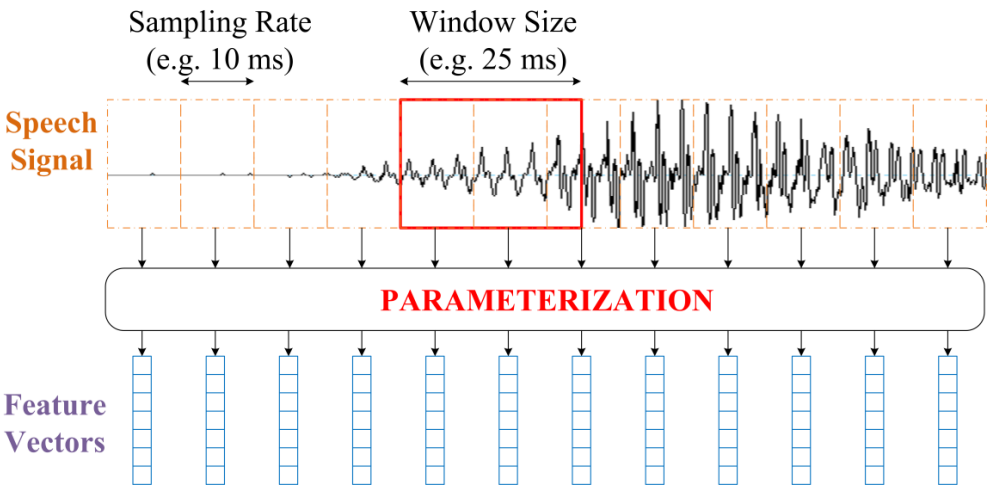


Fig.2.2: Feature vector extraction from speech signal.

To evaluate the performance of ASR system, a common metric called word error rate (**WER**) is used. WER is estimated by aligning the correct word sequence with the recognized word sequence and computing the error rate as,

$$WER = \frac{S + D + I}{N}. \qquad (2.4)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference. Another metric can also be used to evaluate the performance of ASR system is word accuracy (**ACC**)

$$ACC = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N}. \qquad (2.5)$$

where H is the number of correctly recognized words. For some ASR systems which based on syllable, syllable error rate (**SER**) and syllable accuracy (**SACC**) are used instead. In this case, syllable is used in place of word in equations 2.4 and 2.5.

2.2 History of ASR technology

In 1922, a commercial toy named “Radio Rex” was introduced by Elmwood Button Company which probably was the first machine that recognized speech. Rex was a celluloid dog that moved when the spring was released by 500 Hz acoustic energy [17]. But the research of ASR and transcription technology was actually begun in the 1936 at Bell Labs. During 1950s, ASR systems were based on the fundamental ideas of acoustic phonetics. In 1952, an isolated digit recognition system for a single speaker was built at Bell laboratory by Davis, Biddulph, and Balashek [18] which based on measuring or

estimating the formant frequencies during the vowel region of each digit. In 1956, at RCA laboratories, an independent effort by Olson and Belar [19] was performed which tried to recognize 10 distinct syllables of a single speaker, as embodied in 10 monosyllabic words. In 1959, at University College in England, a phoneme recognizer for four vowels and nine consonants was built by Fry [20] and Denes [21]. This recognizer used derivatives of spectral energies as the acoustic information and a simple bigram language model for phonemes to improve the recognition decision. In the same year, at MIT Lincoln laboratories, Forgie and Forgie built a system which was able to recognize 10 vowels in a speaker independent manner [22].

To overcome the issue of computational power of computer, in the 1960s several special purpose hardware were built to perform speech recognition tasks. In 1961, a hardware for vowel recognizer [23] was built by Suzuki and Nakata at the Radio Research Lab in Tokyo. The main component of this system was a filter bank spectrum analyzer whose output from each of the channels was fed in a weighted manner to a vowel decision circuit, and majority decisions logic scheme was used to choose the spoken vowel. In 1962, Sakai and Doshita at Kyoto University developed a hardware for phoneme recognizer [24]. This was the first research on a system that performed speech segmentation along with zero-crossing analysis on different regions of the speech to recognize phonemes. In 1963, at NEC Laboratories, Nagata and coworkers built a hardware for digit recognition [25] which obtained accuracy of 99.7% on 1000 utterances of 20 male speakers. Also in this period, besides segmenting speech, time normalization approach for speech recognition was presented to deal with the non-uniformity of the time scales in speech events. One of the first works was the efforts of Martin and his colleagues at RCA Laboratories in 1964 [26] to develop a set of elementary time normalization methods based on the ability to reliably detect speech starts and ends. These normalization techniques improved the recognition rate by significantly reduce the variability of training and testing material. Another important work at this time is the introduction of dynamic programming or dynamic time warping (DTW) [27] for time aligning of speech utterances by Vintsyuk in the Soviet Union. Although the power of DTW, it was largely unknown in the research community until the late 1970s and early 1980s with works presented in [28-31] that made DTW one of the outstanding methods for speech recognition tasks. Another noticeable achievement in the

1960s was the research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes [32].

In the 1970s speech recognition research obtained some important achievements. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies on the advance use of pattern recognition ideas in speech recognition [33], the power of applying dynamic programming methods [28], and the extension of linear predictive coding (LPC) to speech recognition systems through the use of an appropriate distance measure based on LPC spectral parameters [8]. Another significant achievement was the beginning of large vocabulary speech recognition at IBM Labs in which researchers studied three distinct tasks, namely the New Raleigh language [34] for simple database queries, the laser patent text language [35] for transcribing laser patents, and the office correspondent tasks called Tangora [36], for dictation of simple memos. The third achievement is attempts on speaker independent speech recognition systems at AT&T Bell Labs [37]. In their study, a set of sophisticated clustering algorithms were used to determine the number of distinct patterns required to represent all variations of different words across a wide user population. In the same period, an ambitious ASR project was funded by the Defense Advanced Research Projects Agencies (DARPA) [38] where several speech understanding systems were developed. In 1973 the Heresay-I system introduced by Carnegie Mellon University (CMU) was able to use semantic information to significantly reduce the number of alternatives considered by the recognizer. Also developed by CMU, the Harphy system [39] was shown to be able to recognize speech using a vocabulary of 1,011 words with reasonable accuracy (95% of sentences understood). One particular contribution from the Harpy system was the concept of graph search, where the speech recognition language is represented as a connected network derived from lexical representations of words, with syntactical production rules and word boundary rules. The Harpy system was the first to take advantage of a finite state network (FSN) to reduce computation and efficiently determine the closest matching string. Other systems of note from this project are the CMU's Hearsay-II (pioneered the use of parallel asynchronous processes that simulate the component knowledge sources in a speech system) and BBN's HWIM (incorporated phonological rules to improve phoneme

recognition, handled segmentation ambiguity by a lattice of alternative hypotheses and introduced the concept of word verification at the parametric level).

In the 1980s, researches were concentrated on connected word recognition with the goal of creating a robust system capable of recognizing a fluently spoken string of words. The template-based approach still attracted interest from researchers in this period in which various algorithm based on matching a concatenated pattern of individual words were introduced including the two level dynamic programming approach of Sakoe at Nippon Electric Corporation (NEC) [40], the one pass method of Bridle and Brown at Joint Speech Research Unit (JSRU) in UK [41], the level building approach of Myers and Rabiner at Bell Labs [42], and the frame synchronous level building approach of Lee and Rabiner at Bell Labs [43], etc. Although obtained its own achievement in ASR field, the template-based approach was gradually replaced by statistical approach, especially with the introduction of HMM in the ASR world [9]. The theory of HMM was developed in the late 1960s and early 1970s by Baum, Eagon, Petrie, Soules and Weiss [44, 45] and was applied into speech recognition technology in the 1970s by groups at Carnegie Mellon University and IBM who introduced the use of discrete density HMMs [46-48], and then later at AT&T Bell Laboratories [49-51] where continuous density HMMs were introduced. The main idea was that instead of storing the whole speech pattern in the memory, the units to be recognized are stored as statistical models represented by a finite state automata made of states and links among states. Although the HMM was well known and understood, it was not until widespread publication of the methods and theory of HMMs in the mid-1980s that the technique became widely applied in virtually every speech recognition research in the world. Another noticeable technique re-emerged in the late 1980s was the use of artificial neural networks (ANN) to problems in speech recognition [52]. Several new ways of implementing systems were proposed. For example, a time-delay neural network was used for recognizing consonants [53] and phonemes [54]. But at this time, few researches applied ANN to complex tasks such as large vocabulary continuous speech problems [55]. In 1984, DARPA began a second program to develop a large vocabulary, continuous speech recognition system that yielded high word accuracy for a 1000-word database management task. This program produced a new read speech corpus called Resource Management [56] with 21,000 utterances from 160 speakers, several speech recognition

systems resulted from efforts at CMU with the SPHINX system [57], BBN with the BYBLOS system [58], Lincoln Labs [59], SRI [60], MIT [61], and AT&T Bell Labs [62], and some improvements in the HMM approach for speech recognition tasks.

In the 1990s, a number of innovative error minimization techniques such as discriminative training and kernel-based methods are presented which replaced the traditional frame work of Bayes' problem with an optimization problem involving minimization of the empirical recognition error [63]. This change resulted from the fact that the distribution functions for the speech signal could not be accurately chosen or defined, and the Bayes' decision theory becomes inapplicable under these circumstances. In other words, the ultimate goal of designing a speech recognizer should be to achieve the smallest recognition error rather than obtain the best fitting of a distribution function to the given data as advocated by the Bayes' criterion. An example of discriminative training, the Minimum Classification Error (MCE) criterion was proposed along with a corresponding Generalized Probabilistic Descent (GPD) training algorithm to minimize an objective function which acts to approximate the error rate closely [64]. Another example was the Maximum Mutual Information (MMI) criterion. In MMI training, the mutual information between the acoustic observation and its correct lexical symbol averaged over a training set is maximized. Both the MMI and MCE can lead to speech recognition performance superior to the maximum likelihood based approach [64].

Some important feature transformation techniques were also introduced in the 1990s. First, A new technique for the analysis of speech, called perceptual linear prediction (PLP) technique [65] was presented by Hermansky. This technique used three concepts from psychophysics of hearing to derive an estimate of the auditory spectrum. Other techniques were proposed to alleviate channel distortion and speaker variations like RASTA filtering [66, 67] and Vocal Tract Length Normalization (VTLN) [68, 69], respectively. Also in this period, various methods for adapting the acoustic models to a specific speaker data have been presented. Two commonly used methods are the maximum a posteriori probability (MAP) [70, 71] and the maximum likelihood linear regression (MLLR) [72]. Other methods focused on the HMM training by shifting the paradigm of fitting the HMM to the data distribution to minimizing the recognition error, such as the minimum error discriminative training [73].

HMM-based continuous speech recognition technology, reflecting the computational power of the time, was initial developed in the 1980s which focused on either discrete word speaker dependent large vocabulary systems or whole word small vocabulary speaker independent applications [74]. In the early 1990s, attention switched to continuous speaker independent recognition. Starting with the artificial 1000 word Resource Management task [56], the technology developed rapidly and by the mid-1990s, reasonable accuracy was being achieved for unrestricted speaker independent dictation. Much of this development was driven by a series of DARPA and NSA programs [75]. Many research groups and individuals have contributed to the progress of HMM-based speech recognition in which each will typically have its own architectural perspective. One of the important contributions is the freely available of a speech recognition toolkit named HTK. This is a portable software toolkit for developing system using continuous density hidden Markov models developed by the Cambridge University Speech Group [76].

The success of statistical methods revived the interest from DARPA at the juncture of the 1980s and the 1990s. The DARPA program continued into the 1990s with the read speech program. Following the Resource Management task, the program started another task called the Wall Street Journal [77]. The aim was to recognize read speech from the Wall Street Journal, with a vocabulary size as large as 60,000 words. In parallel, a speech understanding task called Air Travel Information System (ATIS) [78] was developed. The goal of the ATIS task was to perform continuous speech recognition and understanding in the airline reservation domain. The tasks were later expanded to some speech understanding applications including transcription of broadcast news and conversational speech. The Switchboard task is among the most challenging ones proposed by DARPA. In this task, speech is conversational and spontaneous with many instances of so-called disfluencies such as partial words, hesitation and repairs. A number of human language technology projects funded by DARPA in the 1980s and 1990s further accelerated the progress, as evidenced by many papers published in the proceedings of the DARPA Speech and Natural Language/Human Language Workshop.

In the 2000s, different techniques were applied to optimize the speech recognizers. In 2004, A variational Bayesian estimation and clustering (VBEC) techniques were developed [79]. Unlike the maximum likelihood (ML) approach which based on ML parameters, the

total Bayesian framework generates two major Bayesian advantages over the ML approach for the mitigation of over-training effects, as it can select an appropriate model structure without any data set size condition, and can classify categories robustly using a predictive posterior distribution. In 2005, Giuseppe Richardi [80] have developed the technique called Active Learning (AL). The goal of AL is to minimize the human supervision for training acoustic and language models and to maximize the performance given the transcribed and un-transcribed data. With the same purpose of training a recognizer with as little manually transcribed acoustic data as possible, an unsupervised training of acoustic models for large vocabulary continuous speech recognition was proposed in [81]. In [82], they analyzed the differences in acoustic features between spontaneous and read speech using a large-scale spontaneous speech database “Corpus of Spontaneous Japanese (CSJ)” which indicated that spectral reduction is one major reason for the decrease of recognition accuracy of spontaneous speech. Sadaoki Furui [83] investigated SR methods that can adapt to speech variation using a large number of models trained based on clustering techniques. In [84], they explored the application of conditional random fields (CRFs) to combine local posterior estimates provided by multilayer perceptrons (MLPs) corresponding to the frame-level prediction of phone classes and phonological attribute classes. Comparing on phonetic recognition using CRFs to an HMM system trained on the same input features showed that the monophone label CRFs is able to achieve superior performance to a monophone-based HMM and performance comparable to a 16 Gaussian mixture triphone-based HMM. Hermansky proposed a new speech feature that is estimated from an artificial neural net [85]. The features are the posterior probabilities of each possible speech unit estimated from a multi-layer perceptron. Another feature transformation method is feature-space minimum phone error (fMPE) [86]. The fMPE transform operates by projecting from a very high-dimensional, sparse feature space derived from Gaussian posterior probability estimates to the normal recognition feature space, and adding the projected posteriors to the standard features.

The Effective Affordable Reusable Speech-to-Text (EARS) program was conducted to develop speech-to-text (automatic transcription) technology with the aim of achieving substantially richer and much more accurate output than before. The tasks include detection of sentence boundaries, fillers and disfluencies. The program was focusing on natural,

unconstrained human speech from broadcasts and foreign conversational speech in multiple languages. The goal was to make it possible for machines to do a much better job of detecting, extracting, summarizing and translating important information, thus enabling humans to understand what was said by reading transcriptions instead of listening to audio signals [87, 88]. In 2000, the Sphinx group at Carnegie Mellon made available the CMU Sphinx [89], an open-source toolkit for speech recognition.

In summary, there are many noticeable achievements has been obtained in ASR over the period of six decades which make speech recognition become more and more applicable in real-life applications.

Chapter 3

Vietnamese Language and Speech Recognition Studies

3.1 Introduction of Vietnamese

Vietnamese is a Viet-Muong language in the Mon-Khmer group within the Austro-Asiatic family. It is officially used in speaking and writing system of Vietnam [90], and is now spoken as the mother tongue of about 85 million people (including four million abroad). Vietnamese is often erroneously considered to be a monosyllabic language, but it is an isolating language, in which the words are invariable, and syntactic relationships are shown by word order and function words, and so it never changes its morphology. Vietnamese is also a tonal language.

In the phonological form, a Vietnamese word may consist of one or more syllables. Syllable is the smallest unit of pronunciation uttered without interruption, the phonological unit of words. All words are constructed from at least one syllable. Syllable cannot occur by itself unless it is a monosyllabic word. In the Vietnamese lexicon, there are about 80% of words that have two syllables. Some other words have three or four syllables (many of these polysyllabic words are formed by reduplicative derivation).

Additionally, in the morphological form, a Vietnamese word may consist of one or more morphemes. Morpheme is the smallest meaningful unit in the grammar of a language, the morphological unit of word. Being the smallest meaningful element, a morpheme cannot be cut into smaller parts and still retains meaning. While a word can occur freely by itself, a morpheme may or may not be able to. When a morpheme can occur by itself, it is a word with a single morpheme; but when a morpheme cannot occur by itself, it has to be combined with other morphemes to form a word. Poly-morphemic words are either compound words or words consisting of stems plus affixes or resulting from reduplication.

Unlike English, most Vietnamese morphemes consist of only one syllable. Polysyllabic morphemes tend to be borrowed from other languages. Some examples of Vietnamese word are shown in Tab. 3.1. Many of Vietnamese words are created by either compounding or reduplicative derivation. Affixation is a relatively minor derivational process.

Tab. 3.1: Examples of Vietnamese word and their form.

Vietnamese word	English meaning	Morphological form	Phonological form
gạo	rice	Mono-morphemic	monosyllabic
a-xít (loanword)	acid	mono-morphemic	disyllabic
dưa hấu	watermelon	bi-morphemic	disyllabic
điệp điệp trùng trùng	layer upon layer	Poly-morphemic	polysyllabic

In writing system, Vietnamese language uses a set of Latin symbols. It consists of 22 out of 26 letters as in the English alphabet, seven letters which used only in Vietnamese and an addition of five diacritics presenting six tones (Fig.3.1). Moreover, there are nine digraphs (‘ch’, ‘gh’, ‘gi’, ‘kh’, ‘ng’, ‘nh’, ‘ph’, ‘th’, ‘tr’) and one tri-graph (‘ngh’) which are formed from the above letters to present special graphemes of Vietnamese. They used to be considered as independent letters in the old writing system but not in modern Vietnamese. Note that Vietnamese does not have letters ‘f’, ‘j’, ‘w’, and ‘z’ as in English, although they may appear in loanwords or informal writing. In the old Vietnamese writing system, polysyllabic words were written with hyphens to separate the syllables, as in ‘nhà-thờ’ (church), ‘ký-túc-xá’ (dormitory), or ‘cà-phê’ (coffee). Spelling reform proposals have suggested writing these words without spaces (for example, the above words would become ‘nhàthờ’, ‘kýtúcxá’, ‘càphê’). However, the prevailing practice is to omit hyphens and write all polysyllabic words with a space between each syllable.

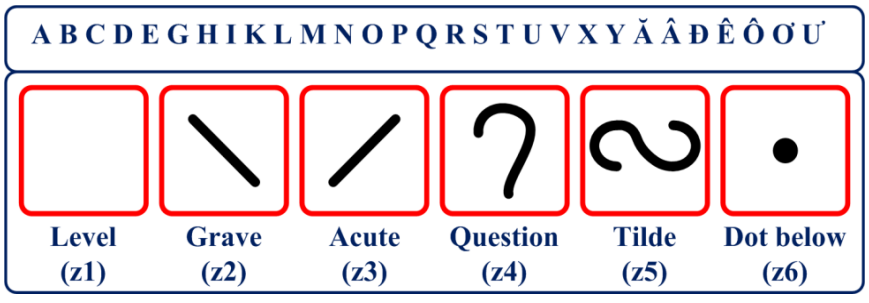


Fig.3.1: The Vietnamese alphabet and tone in writing system.

There are three major dialects of Vietnamese that are geographical in nature: northern, central, and southern. These dialect regions differ mostly in their sound systems, and also in

vocabulary (including basic vocabulary, non-basic vocabulary, and grammatical words), and grammar.

3.2 Vietnamese phonology

3.2.1 Vowels

Comparing to English, Vietnamese has a comparatively large number of vowel phonemes. Tab. 3.2 shows the International Phonetics Association (IPA) chart of monophthongs for northern dialect. Note that, there are three main dialects for Vietnamese language that are geographical in nature: southern, northern and central. Each makes distinctions that the other does not.

Tab. 3.2: IPA chart of monophthongs.

	Front	**Central**	**Back**
Close	i	ɨ	u
Close-mid	e	əː	o
Open-mid	ɛ	ə	ɔ
Open		aː	
		a	

- There are eight unrounded vowels and three back rounded vowels: /u, o, ɔ/
- Vowel /ə/ and /a/ are pronounced shorter than other vowels.
    - Vowel /a/ and /aː/: Short /a/ and long /aː/ are different phonemic vowels, differing in length only (and not quality). (The [ː] symbol indicates a long vowel.)
    - Vowel /ə/ and /əː/: Han [91] suggests that short /ə/ and long /əː/ differ in both height and length, but the difference in length is probably the primary distinction. Thompson [92] seems to suggest that the distinction is due to height (as he does for all Vietnamese vowels), although he also notes the length difference.
- Vowel /ɨ/ is close central unrounded vowel. Many descriptions, such as Thompson, Nguyễn [93], Nguyễn [94], consider this vowel to be close back unrounded vowel: /ɯ/. However, Han's instrumental analysis indicates that it is more central than back. Brunelle [95] and Pham [90] also transcribe this vowel as central.

Besides, There are two semivowel phonemes: /w/ (presented by letter ‘u’, ‘o’) and /j/ (presented by letter ‘y’, ‘i’) in which semivowel /w/ is either in the role of medial sound (like ‘o’ in ‘toán’, ‘toàn’, ‘xoan’, etc., or ‘u’ in ‘tuần’, ‘tuấn’, ‘quẩn’, etc.) or in the role of final sound (like ‘o’ in ‘đào hào’, ‘báo cáo’, etc., or ‘u’ in ‘đau’, ‘rau cau’, etc.), and

semivowel /j/ is in the roll of final-sound. In addition to monophthongs, Vietnamese also has three main diphthongs: /iə/, /ɨə/, /uə/.

In writing system, these vowels are presented using 12 letters. Tab. 3.3 shows the pronunciations and corresponding English sounds of these letters. Note that letters ‘i’ and ‘y’ have the same pronunciation and can be used interchangeably in many Vietnamese syllables, except in syllables with diphthong or triphthong. It is supposed that letter ‘y’ should be used in Sino-Vietnamese syllables (syllables borrowed from Chinese), while letter ‘i’ is for native syllables, but in reality this problem is settled by imitation and habit. Tab. 3.4 and Tab. 3.5 list all possible combinations of two and three vowels occurred in Vietnamese syllables.

Tab. 3.3: Pronunciations of 12 vowel letters.

Spelling	Pronunciation	English approximation
a	/aː/, /a/, /ɜ/	b**a**r
ă	/a/	br**o**ther
â	/ə/	gard**e**n
e	/ɛ/	**e**mbark
ê	/e/, /ə/	m**a**te
i	/i/, /j/	m**i**ni
o	/ɔ/, /aw/, /w/	c**o**rner
ô	/o/, /əw/, /ə/	m**o**bile
ơ	/əː/, /ə/	play**er**
u	/u/, /w/	gl**u**e
ư	/ɨ/	h**u**h
y	/i/, /j/	m**i**ni

Tab. 3.4: Combination of two vowels in Vietnamese.

2 Vowels Combination												
→	**-a**	**-ă**	**-â**	**-e**	**-ê**	**-i**	**-o**	**-ô**	**-ơ**	**-u**	**-ư**	**-y**
a-						ai	ao			au		ay
ă-												
â-										âu		ây
e-							eo					
ê-										êu		
i-	ia				iê					iu		
o-	oa	oă		oe		oi	oo					
ô-						ôi						
ơ-						ơi						
u-	ua	uă	uâ	ue	uê	ui		uô	uơ			uy
ư-	ưa					ưi			ươ	ưu		
y-	ya				yê							

Tab. 3.5: Combination of three vowels in Vietnamese.

3 Vowels Combination										
↓	**iê-**	**oa-**	**oe-**	**ua-**	**uâ-**	**ue-**	**uô-**	**uy-**	**ươ-**	**yê-**
-a								uya		
-ê								uyê		
-i		oai		uai			uôi		ươi	
-o			oeo	uao		ueo				
-u	iêu			uau				uyu	ươu	yêu
-ư										
-y		oay		uay	uây					

3.2.2 Consonants

The set of 23 consonants occurring in the Vietnamese language is shown in Tab. 3.6. An interesting property of Vietnamese consonants is that there are not any consonant groups presented in a syllable as in the English case but each consonant has to follow or precede by a vowel or group of vowels. For example, the English words ‘spring’, ‘kind’, or ‘best’ have three consonant groups ‘spr’, ‘nd’ and ‘st’ respectively, but in Vietnamese there are not such groups of consonants. Remember that dialectal differences exist for Vietnamese language and should be considered when using the phonemic chart of consonants. Not all dialects of Vietnamese have the same consonants in a given syllable (although all dialects use the same spelling in the written system).

All the consonants are presented in the writing system using 17 letters, 9 digraphs and one trigraph as shown in Tab. 3.7. Note that, in some hypotheses, there is another consonant called glottal stop (ʔ) which is not presented by any letter in the writing system. Some letters have more than one pronunciation and are depended on the dialect (southern, northern or central) of the speaker.

Tab. 3.6: IPA chart of Vietnamese consonants.

		Labial	**Alveolar**	**Retroflex**	**Palatal**	**Velar**	**Glottal**
Stop	*voiceless*	p	t	ʈʂ~ʈ	c~tɕ	k	(ʔ)
	aspirated		tʰ				
	voiced	ɓ	ɗ				
Fricative	*voiceless*	f	s	ʂ		x	h
	voiced	v	z	ʐ~ɹ		ɣ	
Nasal		m	n		ɲ	ŋ	
Approximant			l		j		

Tab. 3.7: Pronunciations of consonant letters.

Spelling	Pronunciation	English approximation
b	/b/	**b**ut
c	/k/	**c**ar
d	/z/	**z**oo (Northern)
	/j/	**y**es (Southern & Central)
đ	/ɗ/	**d**o
g	/ɣ/	**g**o
h	/h/	**h**at
k	/k/	**k**ing
l	/l/	**l**ong
m	/m/	**m**e
n	/n/	**n**o
p	/p/	**p**ing **p**ong
q	/k/	**q**ueen
r	/z/	**z**oo (Northern)
	/ʐ/	**r**un (Southern & Central)
s	/s/	**s**tay (Northern)
	/ʂ/	**sh**ow (Southern & Central)
t	/t/	**t**op (unaspirated)
v	/v/	**v**ideo
x	/s/	**s**ee
ch	/c/	**ch**a-**ch**a
gh	/ɣ/	**g**o
gi	/z/	**z**oo (Northern)
	/j/	**y**es (Southern & Central)
kh	/x/	lo**ch**
nh	/ɲ/	ca**ny**on
ng	/ŋ/	si**ng**
ph	/f/	**Ph**ilip
th	/tʰ/	**th**in
tr	/c/	**ch**a-**ch**a (Northern)
	/ʈʂ/	**tr**y (Southern and Central)
ngh	/ŋ/	si**ng**

3.2.3 Tones

Vietnamese can be seen as a tonal language or, more specifically, it utilizes “lexical tone”. This means that the same syllable can have different meanings depending upon the tone with which it is produced. These tones differ from each other in pitch, length, intensity, phonation, and contour melody.

The most popular hypothesis about tone is the six tone system. In this system, there are six tones, five of which presented by marks in the Vietnamese orthography, called diacritical marks or tonal marks. Usually, these tonal marks are added above or under the

vowel letter (monophthong), or the primary vowel letter of a diphthong or a triphthong in a syllable. Tab. 3.8 shows the names, descriptions and corresponding examples of six tones in Vietnamese. These tones are classified into two groups: even and slant tone. Even tone includes *'ngang'* or *'huyền'*; and slant tone includes *'sắc'*, *'hỏi'*, *'ngã'*, *'nặng'*.

Unlike Chinese languages, Vietnamese tones do not rely solely on pitch contour but a combination of phonation type, pitch, length, vowel quality, etc. So perhaps a better description would be that Vietnamese is a register language and not a "pure" tonal language [90]. Fig.3.2 shows the pitch information of six tones in the Hanoi and other northern varieties.

Tab. 3.8: Six Vietnamese tones.

Tone name	Description	Example
ngang (level)	mid level	ba (three)
huyền (grave)	low falling (breathy)	bà (lady)
sắc (acute)	mid rising, tense	bá (governor)
nặng (dot below)	mid falling, glottalized, short	bạ (at random)
hỏi (question)	mid falling(-rising), harsh	bả (poison)
ngã (tilde)	mid rising, glottalized	bã (residue)

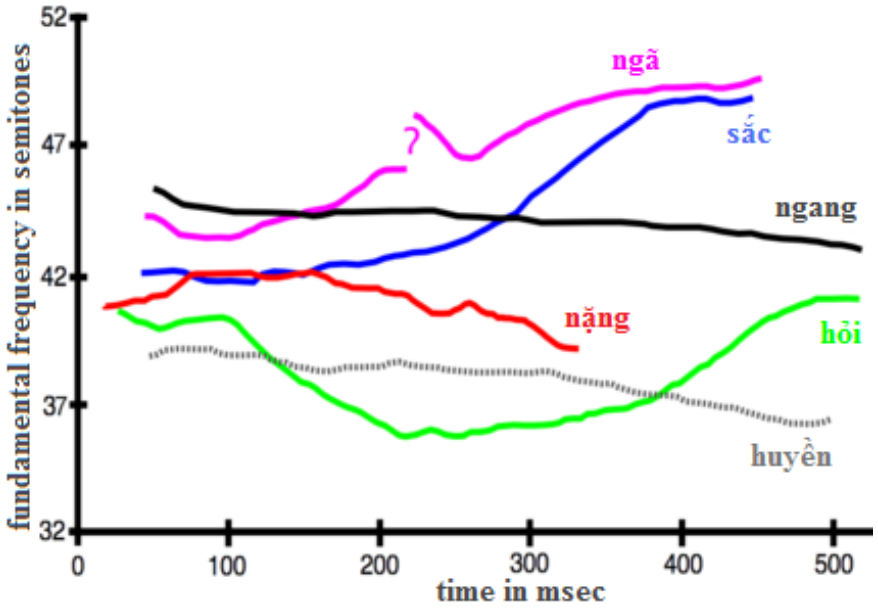


Fig.3.2: Pitch contours and duration of the six Northern Vietnamese tones as uttered by a male speaker (not from Hanoi). Fundamental frequency is plotted over time [96].

Besides, an older analysis assumes that there are eight tones rather than six tones as discussed above. In this eight tone system, two additional tones are presented in Vietnamese for syllables ending in /p/, /t/, /k/ or /c/. This hypothesis is not popular in

modern linguistics because these tones are not phonemically distinct from other tones. Note that in the writing system, only six tones are used.

3.2.4 Syllable

In Vietnamese modern language, there are about 8,000 syllables [97]. Each syllable is modeled as shown in Fig.3.3.

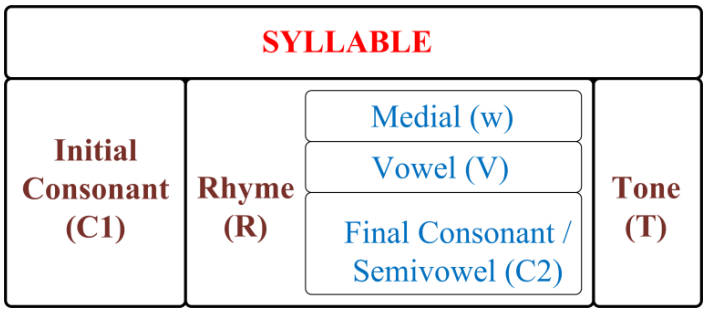


Fig.3.3: Vietnamese syllable's structure.

As described in the figure above, the Vietnamese syllable's structure follows the scheme:

[C1]R + T or **[C1][w]V[C2] + T**

Note that, not all syllables have their complete forms. Some syllables can appear without one of the following components: initial consonant (onset) *C1*, medial *w* and final consonant or semivowel (coda) *C2*. It means main vowel (nucleus) *V* and tone *T* are always presented in the syllable and are the core components of the syllable. Tab. 3.9 shows examples of the syllable's types without tone.

Tab. 3.9: Syllable's type.

Syllable	Example	Syllable	Example
V	á	wV	ọe
VC2	án	wVC2	oán
C1V	ná	C1wV	tóe
C1VC2	nạn	C1wVC2	toán

C1 – The optional onset *C1* is the longest sequence of consonants to the left of the nucleus. For Vietnamese, the onset consists of only one consonant that precedes the vowel in the syllable. Vietnamese has 22 initial consonant phonemes including:

/b, m, f, v, j, t, tʰ, ɗ, n, z, ʐ, s, ʂ, c, ʈ, ɲ, l, k, χ, ŋ, ɣ, h/

The consonant phoneme /p/ occurs as onset only in borrowed words derived from foreign language like French, English, etc., and usually mispronounces as consonant phoneme /b/. Phoneme /p/ in native Vietnamese syllables occurs only as final consonant.

Stemming from Portuguese tradition, Vietnamese letters ‘g’ and ‘ng’ are written differently before front vowel in order to preserve their phonetic value. And so, before vowel letters ‘e’, ‘ê’, ‘i’ and ‘y’, letters ‘g’ and ‘ng’ become ‘gh’ and ‘ngh’ respectively as in syllables ‘ghi’, ‘nghe’, etc. Note that letter pairs ‘g, gh’ and ‘ng, ngh’ present the same sound /ɣ/ and /ŋ/ respectively.

Besides, it is very common for northern speakers to mispronounce between consonant phonemes /l/ and /n/.

W – The optional medial sound /w/, spelled either ‘u’ or ‘o’, indicates labialization or lip-rounding like ‘quả’, ‘thuế’, ‘thủy’, ‘toa’, ‘khỏe’, etc. It does not follow a labial sound like /b/, /m/, /f/, /v/, /w/ except in French loanwords: ‘buýt’, ‘moa’, ‘phuy’, ‘voan’, etc.

The lexemes that have the initial /nw/ are all Sino-Vietnamese words as in ‘noa’ (/nwa/), ‘noãn’ (/nwan/), etc.

Medial sound /w/ cannot be followed by a rounded vowel such as ‘u’, ‘ô’, ‘o’, or ‘uô’. If there is no initial consonant and the vowel nucleus is ‘i’, ‘ê’, ‘yê’, ‘ơ’, or ‘â’, then /w/ is spelled ‘u’ as in ‘*ủy*’ (/wi/), ‘uế’ (/we/), etc. But if the vowel nucleus is ‘a’, ‘ă’, or ‘e’, then /w/ is spelled ‘o’ as in ‘oa’, ‘oă’, ‘oe’.

If the initial consonant is not ‘q’, the same rule is applied, and /w/ is spelled ‘u’ as in ‘tuy’, ‘huế’, ‘thuyền’, ‘khuya’, ‘tuần’ or ‘o’ as in ‘hoa’, ‘khoa’, ‘ngoặc’. If, on the other hand, the syllable starts with ‘q’, then the rhyme sequences /waː/, /wa/, /we/ are spelled ‘ua’, ‘uă’, ‘ue’ like ‘qua’, ‘que’, ‘quặn’, etc.

The sequences /hw/ and /kw/ appears in southern dialect as /w/, excepting spelling pronunciations.

V – The obligatory nucleus that forms the core of a syllable. The vowel nucleus *V* may be any of the following 14 monophthongs or diphthongs: /i/, /ɨ/, /u/, /e/, /əː/, /o/, /ɛ/, /ɜ/, /ɔ/, /a/, /aː/, /uo/, /ie/, /ɨəː/.

C_2 – The optional coda C2 is the final phoneme of a syllable. Vietnamese, like most Austro-Asiatic languages, has fairly restricted phonemic codas including eight consonant

phonemes /c/, /p/, /t/, /k/, /m/, /n/, /ŋ/, /ɲ/ and two semivowels /w/, /j/. These consonants are presented by the following letters and digraphs: ‘c’, ‘m’, ‘n’, ‘p’, ‘t’, ‘ch’, ‘ng’, ‘nh’.

All obstruent codas ‘c’, ‘p’, ‘t’, ‘ch’ are unreleased. English speakers have the habit of releasing their voiceless stops very strongly.

It is not certain what final consonants ‘ch’ and ‘nh’ truly represented at the time the alphabet was made. The pronunciation of these consonants in Vietnamese has had different analyses. One analysis, that of Thompson [92] has them as being phonemes /c/ and /ɲ/, where /c/ contrasts with both final consonants ‘t’ (/t/) and ‘c’ (/k/), and /ɲ/ contrasts with final consonants ‘n’ (/n/) and ‘ng’ (/ŋ/). Final sounds /c/ and /ɲ/ are, then, identified with initial sounds /c/ and /ɲ/. Another analysis has final consonants ‘ch’ and ‘nh’ as representing predictable allophonic variants of the velar phonemes /k/ and /ŋ/ that occur after upper front vowels /i/ (orthographic ‘i’) and /e/ (orthographic ‘ê’). In this work, the first analyzing method is applied.

There are also sound mergers involving final consonants among different regional varieties. Final sounds /t/ and /n/ in northern dialect appear as final sounds /k/ and /ŋ/ in southern dialect, except when these sounds occur after the higher front vocalics /i/, /e/, /j/, in which the southern dialect sounds remain the same as northern dialect sounds /t/ and /n/. Additionally, northern dialect sounds /k/ and /ŋ/ appear as southern dialect sounds /t/ and /n/ when they occur after /i/, /e/, /j/.

T – The tone’s distribution in syllables has a strong relation with the final sound.

Some syllables ending in a glottal stop /p/, /t/ or /k/ only have *‘nặng’* or *‘sắc’* tone. *‘ngang’*, *‘huyền’*, *‘ngã’* and *‘hỏi’* tone cannot exist in these syllables because:

- *‘ngang’* and *‘huyền’* tone have a flat movement. This movement needs a certain length. Meanwhile, some syllables ending in a glottal stop /p/, /t/, /k/ are closed, which make a part of length in the end actually a silent. These two tones, thus, have no opportunity to fully show their flat identity.

- *‘ngã’* tone and *‘hỏi’* tone are broken tones. If they appear in the limited length, they cannot show the complicated movement, either.

In conclusion, of six tones, *‘sắc’* tone and *‘nặng’* tone have the largest distribution in all kinds of syllables.

Note that, southern dialect usually merges *'ngã'* tone and *'hỏi'* tone into one tone resulted in a five tones system in which *'ngã'* tone is removed.

3.3 State of the art ASR of Vietnamese

This section will more concentrate on various perspectives about ASR of Vietnamese than its progression with the aim to emphasize on some basic properties of state of the art ASR of Vietnamese. The first and also one of the most important aspects when doing research on ASR of Vietnamese is the availability of text and speech corpora. For some major well-known languages like English, French, Spanish, etc., the large corpora for research purpose are widely available. But for Vietnamese, an under-resources language, there are not any standard and reliable corpora. The lack of text and speech corpora caused a lot of difficulty for researchers when dealing with ASR of Vietnamese because the fact that collecting of these corpora is a very time-consuming task.

Having a good text corpus is a must for any ASR tasks. At present, the only way to extract a Vietnamese text corpus is to utilize the Internet resources. The text corpora acquired this way can be different in various aspects such as type of websites (literature, news, magazine, entertainment, law, forum, etc.), number of websites for each type, search engine to collect web pages, method to rewrite or normalize special token types (digits, foreign words, abbreviations, acronyms, URL's and e-mail address, etc.) containing in extracted text, method to deal with spelling mistakes occurring in text or how to solve word and sentence segmentation problems.

In [98-102], web robots were used to collect web pages given some starting points on the websites. A large number of websites (about 2500 websites with various categories including daily news, information, entertainment, ecommerce, forum, etc.) which have more pages and richer information than others were collected. The resulted text corpus extracted from the above web pages has the size of 868 MB with 10,020,267 sentences. Tab. 3.10 shows resulted text corpora using four different solutions to filter the extracted text [102]: all-sentences (without sentence filtering); block-based (take only blocks which have at least 5 in-vocabulary words by block), sentence-based (take all sentences containing only in-vocabulary words) and hybrid (take blocks and sentences containing only in-vocabulary words and apply minimal blocks filtering on the rejected sentences). It also shows that text corpora after filtering all the redundant information in the original text

will have a significant reduction in size (54%) but resulted in an improvement of the corpus perplexity by 26%. The corpora in these works can be used for general purpose because they covered lexicon in many fields and the sizes of these corpora are also one of the largest reported corpora.

Tab. 3.10 Text corpus size and perplexity of the language models.

	VN: Web original filter		**VN: Web redundant filter**	
Expe.	*Size (MB)*	*PPL*	*Size (MB)*	*PPL*
all	868	260	402	201
block	667	359	357	282
sent.	370	252	226	195
hybrid	729	259	373	199

In [103-105], the text corpus was collected using the same method presented in [102], but the text data were extracted only from Vietnamese newspapers websites. After removing all superfluous data (menus, references, advertisements, announcements, etc.) presented in the extracted web pages and keeping sentences containing only in-vocabulary syllables (monosyllabic Vietnamese words), the final text corpus has about 2.7 million sentences with 45 million syllables. This corpus is smaller than the above corpus both in size and type of text covered. In [106, 107], the same extracting tool and normalizing methods were used to extract text from the Internet resources. The resulted text corpus contained text mainly in the field of broadcast news with a size of 317 MB and a total 55 million syllables.

In [108-111], text was collected only from newspaper available on the Internet resources. After replacing all numeric expressions, the resulted text corpora contained text with size from 10M to 146M. These text corpora are not too large both in size and information type. In [108], they collected text corpus from only one website – the national newspaper from VOV. On this website, all issues between 6/2004 – 6/2005 were extracted as training data and 200 sentences that were randomly selected from issues between 7/2005 – 8/2005 were used as testing data.

In [112], a text corpus named Vietnamese BachKhoa Text Corpus (BKVTEC) was collected from four large websites in Vietnam which having a size of 535 MB containing 4 million sentences with 90 million syllables. To obtain this text corpus, web pages were filtered and normalized by first creating a dictionary with approximately 6000 Vietnamese

syllables and then extracting and normalizing the text as follows: remove HTML tags, exclude foreign words which are not in the dictionary, replace abbreviations by full words, convert specialcharacters, number, date into text. This corpus is also one of the largest reported text corpora which contained text mainly in two fields: newspapers and literatures, and so it is good for specific purpose. Tab. 3.11 shows some statistic of this corpus.

Tab. 3.11: The size of corpus collected from Vietnamese electronic documents.

Type of text page	Electronic Source	Original size (in HTML format)	Size after filtering	Filtration rate (times)
Electronic Newspapers	Vnexpress.net	2.79 GB	168 MB	16.6
	Vietnamnet.net	1.26 GB	113 MB	11.2
Electronic Literatures	Vanhoc.xitrum.net	415 MB	97 MB	4.3
	Vnthuquan.net	1.2 GB	162 MB	7.4
Totals		4.814 GB	540 MB	8.9

In [113, 114], they used the Rapid Language Adaptation Tools (RLAT) to crawl text from fifteen websites covering mainly Vietnamese newspaper sources with different given link depths. The extracted text was then filtered and normalized using four different steps: remove HTML tags and codes, remove special characters and empty lines, delete lines with less than 75% of tonal syllables (Vietnamese syllables with one of the five diacritics present tones) and delete repeated lines. The resulted text corpus contained roughly 40 million Vietnamese syllables tokens with text only in the field of news. This corpus was used to build language model and select prompts for recording speech data.

Like text corpus, collecting of speech corpus is also an important task related to ASR of Vietnamese. There are three main ways from which speech corpora are collected: (1) collecting audio or video files from Internet resources, (2) recording speech from the selected speakers, and (3) obtaining available speech data from information resources of company, organization, etc. For the first source, speech data of type story reading, news, conversation etc. is collected from audio or video files available on websites. For this type of speech corpus, manually transcribing of speech files is an unavoidable and a very time consuming task. For the second source, speech data is recorded from speakers in controlled condition where the recorded sentences have to be selected to accept some standards or requirements. The third source is not so popular, and usually contains small data for isolated word or small vocabulary speech recognition tasks.

For applications such as telephone number recognition, continuous digit recognition, name recognition, command control, isolated word recognition or connected word recognition, speech data is usually collected using the last two sources. These corpora are usually different in the number of speakers, the vocabulary sizes and the corpus sizes.

In [115, 116], for a name recognition task, the speech data was collected by the immigration control in the International Airport in Ho Chi Minh city from Vietnamese passengers. This data contained 43680 words for training and 4250 words for testing and was divided into three sets: first name, middle names and last name. Each of these words is usually an isolated word.

In [117], a corpus for command control was recorded from10 males and 10 females in which 10 commands corresponding with the fixed positions of robot arm were proposed. Each speaker was asked to utter each command 10 times, 6 times for training and 4 times for testing. This corpus was designed for a speaker dependent speech recognition system in noise condition with four speech versions: clean speech, noisy speech with SNR equals 20, 10 and 5dB.

In [118], a simple Vietnamese corpus for telephone number and name recognition study was proposed. This corpus included 10 speakers (5 males) where each speaker was asked to utter all the telephone numbers from 100 to 199 and some of the names of the telephone. Each sentence was repeated three times and manually labeled using Initial – Final phonetic format. The speech data from 6 speakers (3 males) were used to train the system and speech from the other 4 speakers were used for evaluating purpose.

In [119], they used two speech data for isolated word speech recognition task: the isolated word corpus contained 50 Vietnamese words, each word was spoken 100 times by 5 Vietnamese speakers and the Alphadigit corpus contained a collection of 78044 examples from 3025 speakers uttering six digit strings of letters and digits over the telephone.

In [120], the proposed corpus included 135 isolated words (131 monosyllabic and 4 bi-syllabic words) in which each word contained each of the 16 vowels combined with the 6 tones and had to be number or control command for Internet applications. Each syllable is pronounced 4 times in isolated word mode by 18 speakers (6 females and 2 males from North, 2 females and 2 males from Central, 3 females and 3 males from South). Thus, they

have recorded a set of 9720 words in total for about 3 hours. This corpus was used for tone recognition and isolated word recognition tasks.

In [121], a corpus for continuous digit recognition task was described. This corpus consisted of 442 sentences with 2340 words. It was extracted from two corpora from the Center for Spoken Language Understanding (CSLU): “22 Language v1.2” and “Multi-Language Telephone Speech v1.2”. Each sentence in the corpus consists of digits from 0 to 9. These sentences were recorded from 208 speakers (78 females and 130 males), who recited their telephone numbers, street addresses, ZIP codes or other numeric information over the telephone network in a natural speaking manner. The data were collected from different environments and might contain a noticeable amount of noise and other “real-life” aspects such as breath, glottalization, and music. All the sentences in the corpus have been time-aligned and transcribed at the phonetic level.

On the other hand, with applications specified for LVCSR tasks, text corpora are usually collected using one of the two sources: recorded speech or Internet. For the standard source where speech data is recorded from speakers in some control conditions, there are some popular corpora which were used from many papers. The most popular one was described and used in [98, 100, 101, 103, 104, 107, 122-124]. They built a really complete general purpose Vietnamese speech corpus called the VNSpeechCorpus. In this corpus, there are five different kinds of speech data: (1) Phonemes, (2) Tones, (3) Digits and digit string, (4) application words, and (5) sentences and paragraphs. Phonemes data was created by asking all speakers to utter vowels and consonant solely or in their combined form. For tone data, the same words with 6 different tones were recorded from all speakers. The digit data consisted of isolated digits, connected digits and natural numbers with all of the variants (synonyms) of these digits and numbers. The application words data were also created the same way. This data contains 50 application words where each word corresponds to an action which is useful in several applications such as telephone services, measurement, human-machine interface. For sentence and paragraph data, they first selected the appropriate sentences and paragraphs from text corpus and divided them into common part and private part. The common part was read by all speakers and contained 33 conversations and 37 paragraphs. The private part included about 2000 short paragraphs in which each speaker was asked to read 40 paragraphs. The speech was

recorded in both quiet and office environment. The speakers were from 3 major dialect regions of Vietnam: the South, the North, and the Middle. The age of the speakers range from 15 to 45 years old, among the 50 speakers, 25 speakers are females.

Another noticeable recorded speech corpus was described and used in [114, 125]. This corpus was collected in the GlobalPhone style [126] in which native speakers of Vietnamese were asked to utter prompted sentences of newspaper articles. The speech data was recorded in two phases: In the first phase, speech data was collected from 140 speakers in Hanoi and Ho Chi Minh cities. In the second phase, utterances from the text corpus which cover rare Vietnamese phonemes were first selected. Then 20 Vietnamese speakers who studied in Karlsruhe, German were asked to utter these utterances. All speech data was recorded with a headset microphone in clean conditions. The resulted corpus consists of 25 hours of speech data spoken by 160 speakers. Each speaker uttered about 50 to 200 utterances. Some statistic of the corpus is shown in Tab. 3.12.

Tab. 3.12: Vietnamese GlobalPhone speech corpus.

Set	**#Speaker**		**#Utterances**	**Duration**
	Male	*Female*		
Training	78	62	19,596	22h 15min
Development	6	4	1,291	1h 40min
Evaluation	6	4	1,225	1h 30min
Total	90	70	22,112	25h 25min

In [112], a smaller recorded speech corpus was also described. This speech corpus was created in two steps: first text data was collected from the web with content related to daily life, science, business and car. Then the text data was manually checked and edited resulted in 8047 sentences with about 208000 syllables. These sentences were used to record speech data. The process of data recording was performed in lab with specialized equipment.

The second popular source to collect speech corpora for LVCSR task is from the Internet. This source was widely used in many studies on ASR of Vietnamese. For corpora constructed this way, speech data was collected from websites which are rich of audio or video resource. In [114, 127], a collection of story reading, mailbag, news reports, and colloquy was extracted from the radio program website called “The Voice of Vietnam” (VOV). The resulted corpus contains 23424 utterances obtained from transcription of about 30 male and female broadcasters and visitors. The number of distinct syllables with tone is 4923 and the number of distinct syllables without tone is 2101 in which the data gathered

from the section of story reading is the largest part of the corpus. One deficiency of this corpus is the number of unique speakers is not large enough to cover most variations of Vietnamese speech and the corpus is also not phonetically balanced.

In [108, 110, 111], the speech corpus was also collected from VOV – the national radio broadcaster with a total duration of 20 hours. It was manually transcribed and segmented into sentences, which resulted in a total of 19496 sentences and a vocabulary size of 3174 syllables as shown in Tab. 3.13. In [111], they collected another small speech corpus (14 hours), the AFF Suzuki-cup video database for evaluation purpose. This corpus contains speech from commentators of several football matches in which the transcription text has many special token such as foreign club or player names. The resulted corpus has 11593 sentences with a vocabulary size of 1810 syllables.

Tab. 3.13: The VOV speech corpus.

Dialect	Duration (hours)	# Sentences
Hanoi	18.0	17502
Saigon	2.0	1994
Total	20.0	19496

Other speech corpora of the same category were also presented in some less popular works. In [128], a really large spoken corpus was collected according to some selective criteria such as regions (the North, the Center, the South), sources (broadcast news, telephone, dialogue, monologue, etc.), age ranges (less than 10, from 10 to 20, from 20 to 40, from 40 to 50 and older than 50), and gender (male and female). The corpus contains 60 hours of speech news (51432 sentences, 756348 words) and 30 hours of speech lectures (26891 sentences, 377652 words). In [109], The VNBN speech corpus was collected from February to June 2007, from VOH – the Ho chi minh city broadcaster website, which consisted a total of 27 hours of speech data. They were manually transcribed and segmented into a total of 7,496 sentences with a vocabulary size of 3651 syllables.

Having obtained the text corpus, language model (LM) can be constructed using one of the two methods: syllable-based LM or word-based LM. For the first method, the construction of LM is straightforward because syllable segmentation is simple (syllable is separated by space in Vietnamese writing system) and the vocabulary size is also small (5000 to 8000 syllables). Word-based LM has been reported to be better than syllable-based LM but the construction of this LM type has to base on an error-prone word

segmentation algorithm and the vocabulary size is also bigger. The most difficulty when constructing word-based LM is the segmentation of Vietnamese word which can be done using statistical information, linguistic information or hybrid of both.

Syllable-based LM was used in many studies related to LVCSR of Vietnamese because of its simplicity and small out of vocabulary (OOV) syllable. In [98, 100], syllable-based LMs were build using four different solutions to filter the text corpora. The perplexity values of these LMs on the same test data (Tab. 3.10) showed that preprocessing of raw text corpus will affect the LM. In [99, 101], a vocabulary of 6492 Vietnamese syllables was first collected. Then, a syllable-based statistical trigram LM for Vietnamese was estimated from the text corpus using Katz back off with Good-Turing discounting. It is important to note that in this LM, the unknown syllables are removed since the task is in the framework of closed-vocabulary models. The perplexity value evaluated on the speech test corpus is 109 for Vietnamese.

In [129], Both the bigram and the trigram syllable-based LM were constructed using 10 Mb of text collected from VOV website. A vocabulary of 5000 syllables was selected from syllables occurring at least 12 times in the corpus. An interpolated, trigram LM was estimated by employing a non-linear discounting function and a pruning strategy that deletes trigrams on the basis of their context frequency. Tab. 3.14 shows the perplexity of the bigram and trigram language models on the test set. Obviously, the trigram language model is much better than the bigram. This can be explained that a large number of Vietnamese words contain two syllables. One might think that the bigram LM for Vietnamese is similar to the unigram in English case.

Tab. 3.14: Evaluation of the VOV language models.

	utterances	**var**	**perplexity**
bigram	200	21.0	232.9
trigram	200	25.9	174.2

Tab. 3.15: Evaluation of the language models.

	utterances	**var**	**perplexity**
bigram	135	36.4	212.6
trigram	135	39.9	135.5

In [130], the same method as the previous work was used to build bigram and trigram LM. In particular, a 146 Mb of text collected from newspaper websites was employed. A

vocabulary of 5000 syllables was selected from all syllables occurring in the corpus. Tab. 3.15 shows the perplexity of the bigram and trigram LMs on the test set.

In [111], the same strategy was applied to trained LMs. But in this work, the resulted model was adapted into soccer domain by dynamically modifying model parameters on the basis of what can be extracted from adaptation data. Tab. 3.16 shows the perplexities of resulted LMs on evaluation data. It is obvious that the perplexity of the adapted LM was dramatically reduced ensuring better performance for ASR system.

Tab. 3.16 : Language models perplexities.

	utterances	**perplexity**
bigram	300	237.5
trigram	300	141.8
adapted trigram	300	98.2

In [103-105], syllable-based LM was trained from approximately 2.7 million sentences with 45 million syllables using the CMU SLM toolkit with Good-Turing discounting and Katz back off for smoothing. This LM obtained perplexity of 72.74 on test speech corpus. In [112], they used the same toolkit to train syllable-based LM on data present in Tab. 3.17 with perplexity of bigram and trigram LM are shown in Tab. 3.18.

Tab. 3.17: Text corpus for construction of statistical LM.

Type of text page	**Training Part**	**Test Part**
Newspapers	259 MB	22 MB
Literatures	237.9 MB	21 MB
Newspapers + Literatures	496.9 MB	43 MB

Tab. 3.18: Perplexity of Vietnamese LMs on the test corpus.

Language model (LM)		**Perplexity**	
		Literature test corpus	*Newspaper test corpus*
Bigram LM	*Literature LM*	198.94	362.54
	Newspaper LM	497.41	108.57
	Newspaper + Literature LM	227.94	122.22
Trigram LM	*Literature LM*	131.42	237.24
	Newspaper LM	279.26	65.45
	Newspaper + Literature LM	159.89	62.43

Although syllable-based LM is widely used in many works, word-based LM has also shown many promising aspects for ASR tasks. In [103, 105], word-based LM was also

constructed using two word segmentation methods: use of given semantic vocabulary and use of automatic vocabulary. For the first method, a lexicon of polysyllabic words was given. In the segmentation process, they looked for polysyllabic words from beginning to the end of each sentence, syllable by syllable. In case of ambiguity, the longest grouping was chosen. For the second method, an automatic vocabulary was built using a text corpus of syllables to extract a polysyllabic lexicon. The objective is to extract and to isolate the most frequent sequences of syllables (polysyllabic tokens), whether they have a semantic meaning or not. In that case, token can be a group of syllables, a part of word, a word, either word group, and so, the constructed LM is call multi-syllable-based LM rather than word-based LM. In the first step, the word segmentation of sentence is performed by a dynamic programming algorithm with the syllabic language model (the duration of a token cannot exceed 4 syllables). In the second step, every tokens are analyzed and the most significant tokens are retained (in term of mutual information). The semantic vocabulary has 40342 words and the automatic vocabulary has 39823 tokens.

In [110], Both the trigram syllable-based LM and the multi-syllable-based LM were trained using the SRI LM toolkit [131] with Kneser-Ney smoothing. For the syllable-based LM, a lexicon with the 5000 most frequent syllable was used. The process of building a multi-syllable-based LM consisted of two steps. In the first step, the sentences were segmented into polysyllabic tokens using the maximum match method. A Named-Entity list was also added to the original list at this step to improve segmentation quality. After the segmenting process, a vocabulary was created by select the top 40000 most frequent tokens. All 5000 syllables used for syllable-based LM above are also added to this vocabulary. Tab. 3.19 reports the perplexities of both LMs on the same test set, containing 580 sentences randomly selected from VOV website, issued in 2006.

Tab. 3.19 : Syllable-based and word-based perplexities.

	bigram	**trigram**
Syllable-based LM	188.6	136.2
Multi-syllable-based LM	384.5	321.4

In [113, 114], the same toolkit and method were used to construct both syllable-based and word-based LMs, the differences are only from the text data and the vocabulary size. In [113], they trained two syllable-based LM: a 3-gram LM on the cleaned and normalized text data and another LM on an increased text corpus to improve the performance of the

recognition system. The later was based on the complete training text corpus, enriched by additional vocabulary from the development set. For word-based LM, the word segmentation process was based on the statistical method in which an open source Vietnamese dictionary was presented with about 23000 bi-syllablic words and 6500 monosyllable words. Using crawled text data, the frequencies of all bi-syllablic words were calculated. For each sentence in the text corpus, multisyllabic words were searched syllable by syllable from the beginning to the end of the sentence. Words with higher hit rate than the left and right neighbors were selected as multisyllabic words. In [114], the same method was used to build word-based LM using the cleaned and normalized text data from the extracted text corpus.

The same baseline approach as above can also be seen in works presented in [107, 123, 124]. In these works, syllable-based and word-based trigram LMs were trained from the collected text corpus using the SRILM toolkit with a Good–Turing discounting and Katz back off for smoothing. First, a vocabulary of about 6500 syllables was extracted from a 35000 word vocabulary. Then, to prepare data for word-based LM construction, a basic maximal matching method with dynamic programming was used for word segmentation of text corpus. The perplexity values calculated on the text transcription of speech test set are respectively 109 for syllable-based LM and 201 for word-based LM. In fact, using syllable-based language modeling for Vietnamese reduces the model complexity compared to the word-based language modeling while removing the problem of out of vocabulary (OOV) words or syllables. It is particularly interesting in the case of under-resourced languages where the LM training data are limited. Obviously, the drawback of using syllables instead of words is that for the same order *n*, the n-gram coverage is weaker for syllables.

Although using word-based LM for LVCSR task has proved to be a good approach to deal with Vietnamese language and can be the one for the future, syllable-based LM, at current stage, remains the best approach for many Vietnamese ASR tasks.

One biggest difference between Vietnamese and Western languages such as English, French, etc., is that Vietnamese is a tonal language with six tones. This difference makes speech recognition techniques that validated for these languages could not achieve the same result if they were applied directly on Vietnamese. Tone is the most interesting and also the most difficult problem when dealing with LVCSR of Vietnamese. Generally, there are

three main questions for this problem: What is the role of tone in the relation with other phonetic units? How to integrate tone into the recognizer? And, where is the position of tone in syllable?

For the first question, tone can be considered as a dependent or independent component in a syllable. When using as a dependent component, tone has to be attached to other phonemes to form a phonetic unit. This way the length of tone is equal to the length of the phonetic unit. In [112], four different methods were used to add tone information to syllable. Depending on the method, tone could be attached to main vowel, to rhyme, to the last diphone or to both diphones of a syllable. This way the number of phonetic type will be different for each method. In [132], the tone is integrated to syllable using the same method but in this work 8 tones hypothesis is applied. Also, for recognizer based on phonetic unit or mixture unit, tone is integrated on both main vowel and on final part of syllable. In [113, 114], using the so-called "Explicit tone modeling" scheme, all tonal phonemes (monophthongs, diphthongs, and triphthongs) were modeled with six different models, one per tone. For example, the vowel 'a' is represented by the models 'a1', 'a2', 'a3', 'a4', 'a5', 'a6', where the numerals identify the tones. This scheme results in a total 238 phonetic units. In [111], to better model emotional prosodies, a modification to the acoustic model was proposed, in which tones were integrated into tonal vowels. This resulted in a new acoustic model set consisting of 99 phonetic units (27 phones for consonants, 12 phones for non-tonal vowels, and 60 phones for tonal vowels).

On the other hand, tone can also be used independently as a phoneme in a syllable or as a tone unit in a multi-streams speech recognition system. In [110, 111, 129, 130], tones were treated as distinct phonemes and put immediately after the main sound (vowel). With this approach, the context-dependent model could be built straightforwardly because each syllable will be analyzed into the smallest components (Onset, Medial, Nucleus and Coda) which are the same type as tones. The number of HMMs for modeling each syllable in this method will be significant different. For example, three syllables, namely "ê", "xuyên", "xin" will have 1, 5 and 3 HMMs for modeling them respectively. In [130], they explained that Vietnamese is a monosyllable, tonal language, the duration when uttering each syllable is not significantly different. So having different number of HMM to model each syllable will definitely affect the performance of the recognition system. To overcome this problem,

they proposed a new model which explored the use of both diphthongs and tripthongs as basic units, instead of using just monophthongs as in the previous approach. The new model has totally 77 phonetic units in which 29 units used for representing diphthongs, 8 units for triphthongs, 10 units for monophthong, 23 units for consonants and 5 units for tones . With the new acoustic model, the number of HMMs for each syllable is almost the same. In the above works, tones are modeled in the framework of one stream HMMs where tones (dependent or independent) and other phonetic units are of the same type, and are trained using the same kind of feature vectors. In case of independent tone, by constructing context-dependent HMMs, relationship between tones and other phonetic units can be modeled directly.

For the second question, there are several ways in which tone can be integrated into the recognizer: direct tone integration (feature level integration, model level integration), indirect tone integration or the combination of the methods. When integrating tone in feature level, pitch information in form of fundamental frequency (F0), energy, etc., is extracted from speech signal and added to the current feature vector. Acoustic model training from this feature vector can deal with tone using pitch information incident to it. In [127], F0 features extracted by the Praat tool [133] were concatenated with MFCC features to form a single feature vector. Because F0 contains tone information, this feature vector can be used to model both phonetic units and tone. In [113], to model tones, three methods were used to extract pitch information [134]: (1) using cepstrum, (2) using the root cepstrum, and (3) applying autocorrelation. The cepstrum was computed with a window length of 40ms and the position of the maximum of all cepstral coefficients starting with the 30th coefficient was detected. Furthermore, the position of the three left and right neighbors, and their first and second derivatives were considered. This resulted in 21 coefficients (1 maximum, 3 left neighbor, 3 right neighbor plus the first and second order derivatives) to describe tone.

For model level integration, tones can be modeled in one of the two ways: modeling tones in the same framework with other phonetic units or modeling tones totally independent to other phonetic units in the sense of multi-stream framework. For the first approach, tones are considered as either independent units of the same type with other phonetic units or dependent units that have to attach to other phonemes to form tonal

phonetic units. This approach does not need any special requirements of speech data for tones. But the most important aspect that will affect the performance of the recognizer for this approach is what hypothesis are the tones supposed to be? Namely, how you answer the first and the third question about tone? For the second approach, the tones will be modeled independently to the phonemes in which a tone unit set (in comparison with phonetic unit set) and a tone data are needed to construct a tone recognizer. The advantage of this approach is that the tone recognizer can be applied alone to recognize Vietnamese tones or combine with other recognizers for the purpose of Vietnamese speech recognition. But many difficulties also arise when dealing with this approach. The most difficulty is the need of a tone data. Namely, the boundary of tones in syllable of speech data has to be labeled which is a very difficult and time consuming task, especially for LVCSR applications. Usually, tone is supposed to present on the whole syllable, and so, to create tone corpus, syllable has to be extracted manually or automatically using some syllable boundary detection algorithms. Other difficulties such as the type of feature to present tone or the method to combine tone recognizer with other recognizers are also noticeable.

In [120, 135], a tone recognition problems was addressed. This problem is simplified by using speech data of isolated word (syllable) so that the need for labelling tone boundary in syllable reduce to determining the boundary of syllable which is already known. In [120], the pitch contour extracted from the voice part of the syllable was used to train tone models. Although it is supposed that tone information largely presents on the final part (rhyme) of the syllable, the detection of this part is too difficult for the labeling task. By using feature vector computed using some functions of pitch frequency and energy, different tone recognition results has been obtained for this work. In [135], three different feature types were used to describe tones: fundamental frequency (F0), MFCC and frequency modulation. This work is tended for tone classification task which shows the effect of different feature types on the tone recognizer.

For work presented in [120, 135], syllables are pronounced separately and clearly, therefore, the form of the pitch contour of each tone does not change significantly and almost holds the canonical form. However, in case of continuous speech, the F0 contour of the tone will be affected by many factors such as sentence prosody, tone coarticulation, emotion of speaker, etc. In [103, 104, 136], To integrate the tone model to the acoustic

model in LVCSR task, the tone data was extracted as follows: (1) The syllable boundaries were aligned by using the LIA acoustic modeling toolkit [137]. (2) The syllable boundaries were manually corrected using the Praat tool. (3) The voiced segments of syllables obtained from the previous step were used as tone segments. Fundamental frequency F0 of these tone segments was calculated using the cross-correlation or autocorrelation algorithm and the tone model was trained using these feature vectors. In the decoding stage, the score of a syllable hypothesis will be the sum of weight of model scores (score of LM, score of acoustic model and score on tone model). In [105], the same method was used, but they used MFCC to build tone model instead of F0 features. For indirect tone integration, tones are not modeled in acoustic model but integrated into LM [99-101, 106, 107, 124, 138] or the information about the tone was added to the dictionary in form of a tone tag [113, 114] where these tags were questions to be asked in the context decision tree when building context dependent acoustic models.

For the last question, the position of tone can be supposed to be after [110, 111, 129, 130] or on [111-114] the main vowel, after [129] or on [112, 132] the final component of syllable, on the rhyme [112, 132] of syllable or on the whole syllable[103-105, 114]. In these cases, when tones are put on main vowel, consonant, rhyme, final component or syllable, they are considered as dependent component and have to be attached to other components of a syllable to form a phonetic unit, and so, the tones will have the same duration with these components. In other cases, tones are considered as independent phonetic units with the same type as other components of syllable.

Another aspect that has great effect on ASR of Vietnamese is the type of phonetic unit used for constructing acoustic model. The selection of phonetic unit is not straightforward as in English case where each phoneme is a phonetic unit but depends on the analysis of Vietnamese syllable and the knowledge of Vietnamese phonology. Based on the structure of Vietnamese syllable (Fig.3.3) there are two main phonetic unit types: phoneme-based and rhyme-based phonetic unit. In the first type [99, 110-112, 124, 129, 130, 132], the smallest components of syllable (Onset, Medial, Nucleus and Coda) are used as phonetic units, and in the second type [112, 132], the phonetic units are Onset (Initial) and Rhyme (Final) of the syllable. Another type was also presented in [112] where each syllable was analyzed into two sets of phones: /Onset/-/Medial/-/Nucleus/ and /Nucleus/-/Coda/. These

types of phonetic unit are different to one another in the properties of phonetic unit obtained from each syllable and in the total number of distinct phonetic unit.

In addition, using the knowledge of Vietnamese phonology, other types of phonetic unit were also proposed. In [132], the phonetic units were the mixture of units from two types: the first type contains syllables analyzed into two parts: Initial and Rhyme (Rhyme is the combination of two neighbors short phonetic units) and the second type is the other syllables which consist of four phonetic units (Onset, Medial, Nucleus and Coda). The way in which syllables are classified into type 1 or 2 is totally based on the knowledge of Vietnamese phonology. In [113, 114], the syllable was analyzed into initial consonant (Onset) placed at the start of syllable, final consonant (Coda) placed at the end of syllable, and main vowel (monophthong, diphthong or triphthong) which were then used as basic phonetic units. The total number of phonetic unit is 58 phonemes (22 consonants, 11 vowels, 21 diphthongs and 4 triphthongs) when using “Data-driven tone modeling” method and 238 phonemes when using “Explicit tone modeling” method. In [130], a similar approach as in [113, 114] was proposed to obtain a new phonetic unit set which explored the use of both diphthongs and triphthongs as basic units, instead of using just monophthongs. The basic idea of this approach is to overcome the problem of using different number of HMMs for modeling each syllable. This new phonetic unit set has totally 77 phonetic units in which 29 units used for representing diphthongs, 8 units for triphthongs, 10 units for monophthongs, 23 units for consonants and 5 units for tones. Note that, in these methods, the way in which tone is integrated into the syllable will affect the number of distinct phonetic unit as well as the properties of each phonetic unit. Usually, in these works, when tone is used as independent phonetic unit, the number of distinct phonetic unit is smaller than the case where tone is used as dependent one.

The knowledge of Vietnamese phonology, namely what hypothesis is used to explain the phonology relationship, also affects the phonetic unit set. The most obvious different in these hypotheses is the proposal of different set of diphthong, triphthong and even monophthongs. This can be shown in work presented in [113, 114] and [130] where the first work proposed a set of 58 phonetic units and the second one proposed a set of 77 phonetic units. The biggest problem of all the works related to ASR of Vietnamese is that the phonetic unit set is not presented obviously in their works except some grapheme-based

phonetic unit set where each grapheme is used as a phonetic unit. This make works on ASR of Vietnamese become more and more difficult to follow, compare or inherit.

Another difficulty when doing experiments on LVCSR of Vietnamese is the lack of a standard pronunciation dictionary. In many works [99, 110-112, 124, 129, 130, 132], the pronunciation dictionary was created by simply splitting the dictionary entries (syllable, word) into its corresponding graphemes, and so it is called the grapheme-based pronunciation dictionary. For this kind of dictionary, the phonetic units are formed from a single grapheme or a group of graphemes depending on the method or hypothesis used. In some other works, phoneme-based pronunciation dictionary is used in which a grapheme-to-phoneme mapping table of the dictionary entries is introduced. In [113], pronunciation dictionary was built by grapheme-to-phoneme mapping, but phoneme set is not given. In [114], the RLAT tools was used to generate the dictionary which enable the user to learn pronunciation rules by providing initial letter-to-sound mappings and interactively confirming or correcting pronunciation entries. The resulted dictionary was also modeled to present the impact of dialectal variations by using multiple pronunciations for a dictionary entry, but again, the phoneme set was not proposed. In [98], an IPA reference table for sub-word units was used to phonologically transcribe the vocabulary's entries into their corresponding phonetic units, and a tool called VNPhoneAnalyzer was also constructed for creating pronunciation dictionary using the proposed table. This tool was used in many work [100, 101, 103, 104, 107, 122, 123] related to ASR of Vietnamese.

Chapter 4

Thesis Goals

The research on automatic speech recognition of Vietnamese has made large progress in recent years. However, ASR of Vietnamese is just at its experimental stage and yet to reach the performance level required to be widely used in real-life applications. This thesis presents the work on automatic speech recognition of Vietnamese, a tonal, syllable-based language in which new approaches have to be applied to obtain reliable results. When dealing with Vietnamese, the following basic problems have to be solved or clarified: the selection of phonetic unit to build acoustic models, the collection of text and speech corpora, the creation of pronunciation dictionary, the construction of language model and especially, the methods to deal with tone. Another problem which also attracts much interest is the performance of speech recognition systems that are to operate in noisy conditions and that are equipped also with visual information.

With the basic idea of systematically and methodically finding solutions to all the problems mention above, in this work, the following tasks have to be completed:

1. Proposal of a phoneme set optimal to Vietnamese
2. Creation of pronunciation dictionary for Vietnamese
3. Investigation focused on the optimal way to integrate tone in ASR of Vietnamese
4. Proposal of strategies to deal with LVCSR of Vietnamese
5. Preparation of text and speech corpus to train acoustic and language models for Vietnamese
6. Study on various types of language model suited for specific needs of Vietnamese
7. Evaluation of the proposed language model, tone’s hypotheses, and speech recognition strategies on the collected speech corpus
8. Improving speech recognition in noisy condition

Chapter 5

Strategies for Speech Recognition of Vietnamese

5.1 Phoneme set proposal

For experiments on continuous speech recognition, there have to be a set of standard phoneme which will be used to create a pronunciation dictionary. For example, the CMU pronunciation dictionary contains a set of 39 standard phonemes that were used in continuous speech recognition of English (Tab. 5.1).

Tab. 5.1: 39 English Phonemes in CMU pronunciation dictionary.

AA	AE	AH	AO	AW	AY	B
CH	D	DH	EH	ER	EY	F
G	HH	IH	IY	JH	K	L
M	N	NG	OW	OY	P	R
S	SH	T	TH	UH	UW	V
W	Y	Z	ZH			

The need for a set of standard phoneme is obvious but the problem is more complicated in the case of Vietnamese language. From the structure of Vietnamese syllable (see Fig.3.3), it can be seen that the problem here is the presence of tone in the syllable. Tone is the first and also the most difficult aspect that one has to solve when doing experiments on LVCSR of Vietnamese. With the presence of tone, phonemes are not always the basic units to train acoustic models for recognizers but the phonetic units are used instead. Note that, phonetic unit can be the phoneme itself or the combination of phoneme or set of phonemes with tone, which will be clarified in the next section. The way in which a syllable is analyzed into smaller components and the way a tone is integrated into syllable will affect the set of standard phonetic unit.

Tab. 5.2: Grapheme-to-phoneme mapping table.

Type	IPA	Grapheme	Phoneme	Properties
Consonant	ɓ	b	B	voice bilabial implosive
	c	ch	CH	voiceless palatal stop
	ɗ	đ	D	stop, voiced alveolar implosive
	f	ph	F	voiceless labiodental fricative
	h	h	H	voiceless glottal fricative
	j	d	Y	palatal approximant
	k	k, q, c	K	voiceless velar stop
	l	l	L	alveolar lateral approximant
	m	m	M	bilabial nasal
	n	n	N	alveolar nasal
	ɲ	nh	NH	palatal nasal
	ŋ	ng, ngh	NG	velar nasal
	p	p	P	voiceless bilabial stop
	s	x	S	voiceless alveolar sibilant
	ʂ	s	SH	voiceless retroflex sibilant
	t	t	T	voiceless alveolar stop
	tʰ	th	TH	stop, aspirated, alveolar
	ʈʂ	tr	TR	voiceless retroflex affricate
	v	v	V	voiced labiodentals fricative
	x	kh	KH	voiceless velar fricative
	ɣ	g, gh	G	voiced velar fricative
	z	gi	ZH	voiced alveolar sibilant
	ʐ	r	R	voiced retroflex sibilant
Medial / semivowel	w	u, o	W	velar glide
Semivowel	j	y, i	IH	palatal glide
Monophthong	a	a, ă	AU	open front unrounded
	e	ê	EE	close-mid front unrounded
	ɛ	e	EH	open-mid front unrounded
	ə	â	AH	mid-central
	əː	ơ	AX	lower mid, central
	i	y, i	IY	close front unrounded
	ɨ	ư	UH	close central unrounded
	o	ô	AO	close-mid back rounded
	u	u	UW	close back rounded
	ɔ	oo, o	OO	open-mid back rounded
	aː	a	AA	open unrounded
Diphthong	iə	yê, iê, ya, ia	IE	
	ɨə	ươ, ưa	UA	
	uə	uô, ua	UO	

As described in the previous chapter, ASR of Vietnamese is based on two basic phoneme set types: phone-based phoneme set and grapheme-based phoneme set. In many of the previous works, grapheme-based phoneme set was used because of its simplicity: the phoneme is the grapheme itself. Tab. 5.2 shows all graphemes for Vietnamese. The

problem of this phoneme set type is that there are some cases in Vietnamese where two or more different graphemes are pronounced the same. For example, the grapheme pairs “g, gh” and “ng, ngh” both present the same phones /ɣ/ and /ŋ/ respectively but are used separately as phonemes. To solve this problem, in this work a standard phone-based phoneme set is proposed and a many to one grapheme-to-phoneme mapping table is presented. This Vietnamese phoneme set consists of 23 consonants, 11 monophthongs, 3 diphthongs, one medial phoneme and two semivowel phonemes. Tab. 5.2 shows the complete phoneme set with their corresponding graphemes, IPA representation and properties. Note that IPA representation and properties of some phonemes may vary depending on the dialect.

5.2 Pronunciation dictionary creation

The creation of a pronunciation dictionary is one of the most difficult tasks when dealing with Vietnamese language. This problem arises from the lack of a standard pronunciation dictionary for ASR of Vietnamese. When constructing a pronunciation dictionary, the type of resulted dictionary is depended on the selection of phoneme set, the analysis of Vietnamese syllable, and the integration of tone into syllable. As mentioned above, there are two types of phoneme set, and so, there are also two types of pronunciation dictionary including phone-based and grapheme-based type. From the structure of Vietnamese syllable (Fig.3.3), there are several ways in which a syllable can be analyzed into smaller components. Tab. 5.3 shows the four schemes that are used to analyze a Vietnamese syllable. In these schemes, tone is removed from the resulted components because it will be integrated later into the syllable based on hypotheses about tone. For the phoneme-based scheme, a syllable will be analyzed into four components in which each component corresponds with a phoneme. Note that, in these components, only main vowel *V* is obligatory, other components (components inside the square brackets) are optional. In the vowel-based scheme, three components extracted from the syllable are the optional initial phoneme *C1*, the obligatory vowel (monophthong, diphthong or triphthong) *M = [w]V1,* and the optional final phoneme *C*. Vowel *V1* can be the main vowel *V* or the combination of main vowel *V* with the semivowel part of *C2* (note that *C2* can be semivowels or consonants), and *C* is the consonant part of *C2* without semivowel (Section 3.2.4). With the rhyme-based scheme, each syllable consists of two components: initial consonant *C1*

corresponding to a phoneme and rhyme *R = [w]V[C2]* corresponding to a combination of three phonemes. The resulted component from the syllable-based scheme is actually the syllable itself without tone.

Tab. 5.3: Vietnamese syllable analyzing schemes.

Analyzing scheme	**Basic components**
Phoneme-based	[C1] + [w] + V + [C2]
Vowel-based	[C1] + M + [C]
Rhyme-based	[C1] + R
Syllable-based	S

The integration of tone into syllable is depended on the hypotheses about tone, namely, whether the tone is used or not, what is the role of tone in syllable (dependent or independent tone), and where is tone located in syllable. Tab. 5.4 shows various pronunciation dictionary types which are created using different schemes of analyzing a Vietnamese syllable and different hypotheses about tone. One obvious difference between these pronunciation dictionaries is the number of basic phonetic units (see Tab. 5.5). For the same pronunciation dictionary type, grapheme-based dictionary has larger number of phonetic units than the phone-based one. This is because the number of graphemes is larger than the number of phonemes (Tab. 5.2). It can also be seen that, for the same pronunciation dictionary type, dictionary with dependent tone has larger number of basic phonetic units than the one with independent tone. The reason is, for pronunciation dictionaries with independent tone, there are only six tonal phonetic units that present six tones (‘z1’, ‘z2’, ‘z3’, ‘z4’, ‘z5’, ‘z6’) and the other phonetic units do not carry tone information. On the other hand, for pronunciation dictionaries with dependent tone, every phonetic unit that has tone attached to it can produce one of the six different tonal phonetic units. For example phonetic unit ‘o’ with tonal information will have six corresponding tonal phonetic units: ‘oz1’, ‘oz2’, ‘oz3’, ‘oz4’, ‘oz5’, ‘oz6’. Tab. 5.5 also shows that pronunciation dictionaries with the same hypotheses about tone (tone’s role and position) will have different number of basic phonetic units depended on what scheme is used to analyze syllables. Usually syllable-based scheme has the largest number phonetic unit, then the rhyme-based scheme, the vowel-based scheme and finally the phoneme-based scheme. For example, dictionary of type *ST_D* can have up to 20000 units, *C1RT_D* has 641 units, *C1MC2T_D* has 303 units, and *C1wVC2T_D* has 153 units. It is easy to see that scheme

whose component resulted from the combination of more phonemes has larger number of basic phonetic units. Another difference between these pronunciation dictionaries is the property of the phonetic units which is depended mainly on hypotheses about tone. An example of this is the case of phoneme-based analyzing scheme. Three pronunciation dictionaries of type *C1wVC2T_I, C1wVTC2_I* and *C1wVTC2T_I* all have the same basic phonetic units but the property of these units is different. The same thing can also be seen in pronunciation dictionaries of type *C1MCT_I* and *C1MTC_I* of the vowel-based analyzing scheme. The detail of these differences will be explained in the next section.

Tab. 5.4: Vietnamese pronunciation dictionary types.

Tone hypothesis	**Syllable analyzing scheme**			
	Phoneme	*Vowel*	*Rhyme*	*Syllable*
No Tone	C1wVC2	C1MC	C1R	S
Dependent Tone at the end of syllable	C1wVC2T_D	C1MCT_D	C1RT_D	
Independent Tone at the end of syllable	C1wVC2T_I	C1MCT_I	C1RT_I	ST_I
Dependent Tone after (main) vowel	C1wVTC2_D	C1MTC_D		
Independent Tone after (main) vowel	C1wVTC2_I	C1MTC_I		
Dependent Tone present both after main vowel and at the end of syllable	C1wVTC2T_D			
Independent Tone present both after main vowel and at the end of syllable	C1wVTC2T_I			
Dependent Tone on the whole syllable				ST_D

Tab. 5.5: Number of possible basic phonetic unit.

Dictionary type	**Number of phonetic unit**	
	Phone-based	*Grapheme-based*
C1wVC2	39	48
C1wVC2T_D	153	154
C1wVC2T_I	45	54
C1wVTC2_D	109	149
C1wVTC2_I	45	54
C1wVTC2T_D	151	178
C1wVTC2T_I	45	54
C1MC	72	90
C1MCT_D	303	341
C1MCT_I	78	96
C1MTC_D	281	345
C1MTC_I	78	96
C1R	178	208
C1RT_D	641	715
C1RT_I	184	214
S	From 2000 to 4000	From 2000 to 4000
ST_D	From 7000 to 20000	From 7000 to 20000
ST_I	From 2000 to 4000	From 2000 to 4000

To create phone-based or grapheme-based pronunciation dictionary, the grapheme-to-phoneme mapping table and the schemes of analyzing of Vietnamese syllable are utilized. Each entry of the dictionary is created using one of the two procedures described below:

Procedure 5.1: Grapheme-based dictionary entry creation

1. Separate dictionary entry into syllables.
2. Select the first syllable and analyze it into four components: *C1*, *w*, *V* and *C2*
3. Combine phonemes into phonetic unit using one of the syllable analyzing schemes.
4. Integrate tone *T* into syllable to create appropriate tonal phonetic units based on hypotheses about role and position of tone.
5. Concatenate these phonetic units to those of the previous syllables (if any) to create the pronouncing representation of the entry.
6. Select the next syllable (if any), analyze it into four components: *C1*, *w*, *V* and *C2* and go to step 3.

Procedure 5.2: Phone-based dictionary entry creation

1. Separate dictionary entry into syllables.
2. Select the first syllable and analyze it into four components: *C1*, *w*, *V* and *C2*
3. Using mapping table to convert four components *C1*, *w*, *V* and *C2* into their corresponding phonemes.
4. Combine phonemes into phonetic unit using one of the syllable analyzing schemes.
5. Integrate tone *T* into phonemes to create appropriate phonetic units based on hypotheses about role and position of tone.
6. Concatenate these phonetic units to those of the previous syllables (if any) to create the pronouncing representation of the entry.
7. Select the next syllable (if any), analyze it into four components: C1, w, V and C2 and go to step 3.

5.3 Strategies for speech recognition of Vietnamese

5.3.1 Phoneme-based strategy

For this strategy, the basic phonetic units to build acoustic models are phonemes including initial consonant *C1*, medial *w*, main vowel *V* and final phoneme *C2*. This strategy consists of 7 basic methods for each of the phone-based and grapheme-based phoneme set. In these

methods, methods with dependent tone always have larger number of basic phonetic units than methods with independent tone or without using tone (Tab. 5.5). Note that, the number of basic phonetic units of these methods is also smaller than the corresponding methods (methods with the same tone hypotheses) of other strategies. The most advantage of this strategy is acoustic models can be built using the same strategy applied for ASR of English. But the most difficulty is the proposal of appropriate hypotheses about tone. Fig.5.1 shows the tone and pitch information when analyzing syllable 'TOÁN' using seven methods of the phoneme-based strategy.

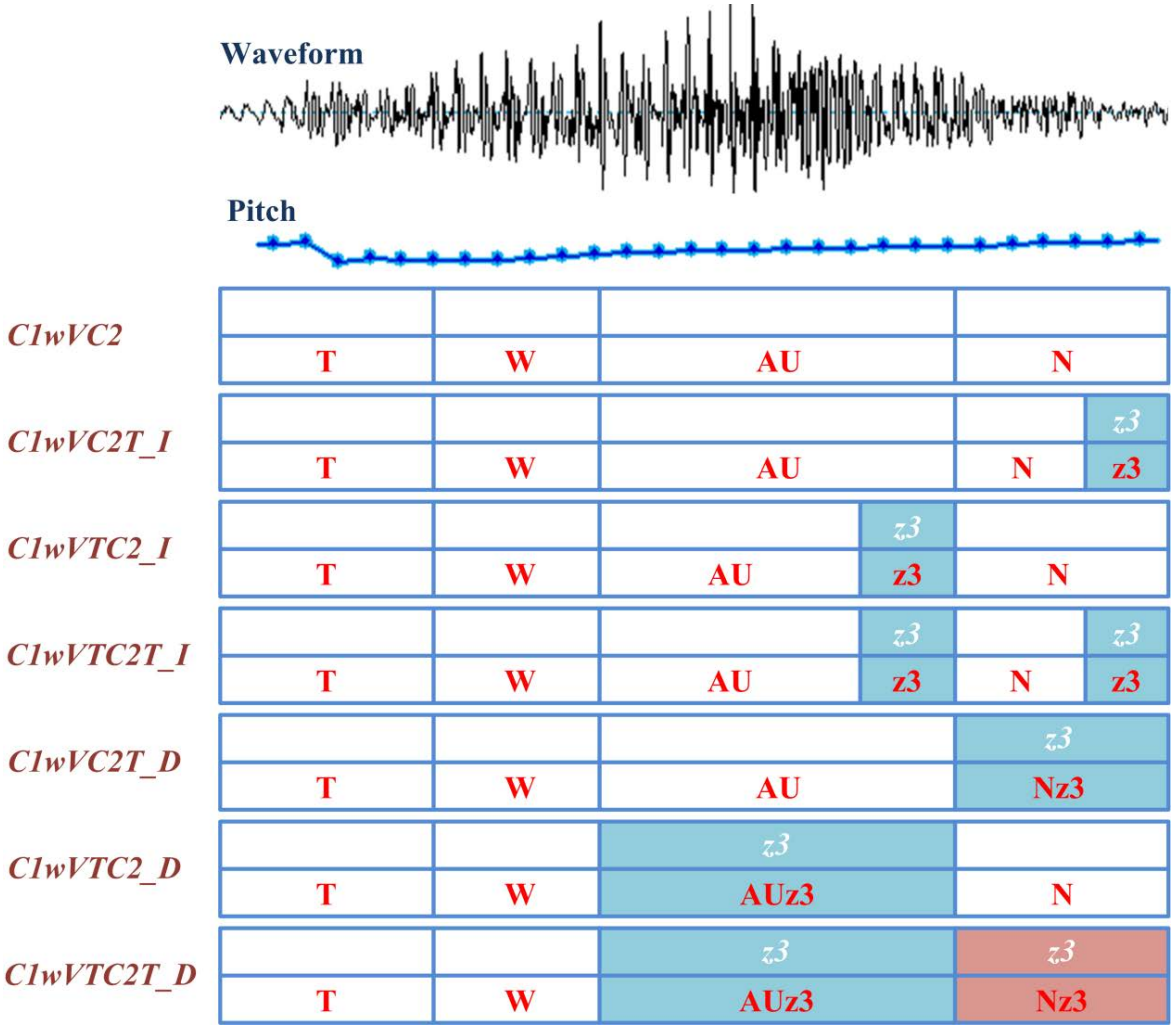


Fig.5.1: Analysis of syllable 'TOÁN' in phoneme-based strategy.

5.3.1.1 Without tone analysis: C1wVC2

For this method, a syllable can be analyzed into at most four phonemes (*C1*, *w*, *V*, *C2*) in which tone information is not integrated into these phonemes resulted in a total 39 or 48

basic phonetic units depended on the phoneme set type. When building context-independent or context-dependent acoustic model, the same method can be applied as in the English case where the number of phonetic units of this method is as small as the English phoneme set (see Tab. A.1 and Tab. A.2 in the appendix). The biggest drawback of this method is that there is not any tone information incident to any of the phonetic units and so it will affect the final result of speech recognition.

Tab. 5.6 shows an example of analyzing the Vietnamese syllable 'TOAN' into its corresponding phonetic units. This syllable with six tones will form six different syllables in Vietnamese language but these syllables are presented by only one pronunciation dictionary entry. Fig.5.1 presents the visual information of syllable 'TOÁN' resulted from this method. The first band of the method *C1wVC2* shows the state of tone, and the second band shows all the phonetic units of the syllable.

Tab. 5.6: Analyzing of syllable 'TOAN' using C1wVC2 method.

TOAN		**Phone-based pronunciation dictionary entry**
Tone	*Syllable*	*C1wVC2*
z1	TOAN	T W AU N
z2	TOÀN	
z3	TOÁN	
z4	TOẠN	
z5	TOẢN	
z6	TOÃN	

5.3.1.2 Independent tone analysis: C1wVC2T_I, C1wVTC2_I and C1wVTC2T_I

For this analyzing approach, tone is considered as an independent phonetic unit with the same role as other phonemes in a syllable. This group of analysis contains three methods (*C1wVC2T_I, C1wVTC2_I* and *C1wVTC2T_I*) which share a similar property: the pronunciation dictionaries are based on the same set of basic phonetic units (see Tab. A.3 and Tab. A.4). In these methods, the four phonemes *C1, w, V, C2* extracted from a syllable are considered as phonetic units without tone information and another tone phoneme which is put into the syllable after main vowel *V* or after the last phoneme *C2* is the only tonal phonetic unit of this syllable. Because there are six tones, each of the three methods will have only six tonal phonetic units: 'z1', 'z2', 'z3', 'z4', 'z5', 'z6'. This makes the basic phonetic unit set of these methods are nearly the same as the method *C1wVC2* with a difference is only from the addition of six tonal phonetic units.

Although having the same basic phonetic unit set, these methods are different from each other in the position of tone in syllable as well as the property of tone. For the method *C1wVC2T_I*, tone is the last component which puts after main vowel *V* or final phoneme *C2* in a syllable and is a tonal phonetic unit. Note that tone is put after the main vowel only in syllables without final phoneme *C2*. In this method, tone captures only the last part of pitch information incident to the syllable. For the method *C1wVTC2_I*, tone is put after the main vowel *V* of a syllable, and so, it occupies only the pitch information part between main vowel *V* and final phoneme *C2* (Fig.5.1). On the other hand, the method *C1wVTC2T_I* tends to put the same tone after both of the main vowel *V* and the final phoneme *C2* to form two identical tonal phonetic units. Note that some syllables do not have the phoneme *C2* so there will be only one tonal phonetic unit in this case. The problem of this method is that two identical tonal phonetic units are needed to capture two different parts of pitch information incident to a syllable. Tab. 5.7 shows some differences in the analysis of syllable 'TOAN' among these methods.

Tab. 5.7: Analyzing of syllable 'TOAN' using C1wVC2T_I, C1wVTC2_I and C1wVTC2T_I methods.

TOAN		**Phone-based pronunciation dictionary entry**		
Tone	*Syllable*	*C1wVC2T*_I	*C1wVTC2*_I	*C1wVTC2T_I*
z1	TOAN	T W AU N z1	T W AU z1 N	T W AU z1 N z1
z2	TOÀN	T W AU N z2	T W AU z2 N	T W AU z2 N z2
z3	TOÁN	T W AU N z3	T W AU z3 N	T W AU z3 N z3
z4	TOẠN	T W AU N z4	T W AU z4 N	T W AU z4 N z4
z5	TOẢN	T W AU N z5	T W AU z5 N	T W AU z5 N z5
z6	TOÃN	T W AU N z6	T W AU z6 N	T W AU z6 N z6

5.3.1.3 Dependent tone analysis: C1wVC2T_D, C1wVTC2_D and C1wVTC2T_D

Three methods *C1wVC2T_D*, *C1wVTC2_D* and *C1wVTC2T_D* are grouped into this category. The detail of the phonetic unit sets is shown in Tab. A.5 to Tab. A.10. For these methods, a syllable can be analyzed into at most four phonemes and the tone which is considered as a dependent component will be attached to at least one of these phonemes to form a tonal phonetic unit. The remaining phonemes will form phonetic units without tone information. The most difference between these methods and the previous methods is the presence of many tonal phonetic unit types in which one phoneme with tone will present six different tonal phonetic units. For example, phoneme 'a' with tone will have one of the six tonal phonetic units 'az1', 'az2', 'az3', 'az4', 'az5', 'az6' related to it. This makes the

number of total basic phonetic unit of these methods are always larger than other methods in the same phoneme-based strategy. Another noticeable property is that tone will have the same length (time boundary) with the phoneme it is attached to (Fig.5.1).

For the method *C1wVC2T_D*, tone is attached to the last part (main vowel *V* or final phoneme *C2*) of the syllable to form a tonal phonetic unit. Note that tone is attached to the main vowel only in syllable without final component *C2*. In this method, tone captures only the last part of pitch information in the form of dependent tone attached to a phoneme of the syllable. For the method *C1wVTC2_D*, tone is attached to the main vowel *V* of a syllable only, and so the number of distinct phonetic unit will be less than in comparison with method *C1wVC2T_D*. In this method, only the vowel part of speech signal carries tone information. On the other hand, the method *C1wVTC2T_D* tends to attach the same tone to both of the main vowel *V* and the final phoneme *C2* at the same time to form at the most two different tonal phonetic units (note that some syllables do not have the component C2). This method supposes that two tonal phonetic units are needed to capture the tone information incident to a syllable. Tab. 5.8 shows some differences in the analysis of syllable 'TOAN' among these methods.

Tab. 5.8: Analyzing of syllable 'TOAN' using C1wVC2T_D, C1wVTC2_D and C1wVTC2T_D methods.

TOAN		**Phone-based pronunciation dictionary entry**		
Tone	*Syllable*	*C1wVC2T_D*	*C1wVTC2_D*	*C1wVTC2T_D*
z1	TOAN	T W AU Nz1	T W AUz1 N	T W AUz1 Nz1
z2	TOÀN	T W AU Nz2	T W AUz2 N	T W AUz2 Nz2
z3	TOÁN	T W AU Nz3	T W AUz3 N	T W AUz3 Nz3
z4	TOẠN	T W AU Nz4	T W AUz4 N	T W AUz4 Nz4
z5	TOẢN	T W AU Nz5	T W AUz5 N	T W AUz5 Nz5
z6	TOÃN	T W AU Nz6	T W AUz6 N	T W AUz6 Nz6

5.3.2 Vowel-based strategy

For this strategy, the basic phonetic units to build acoustic models are initial consonant *C1*, vowel *M* and final consonant *C*. This strategy consists of five basic methods for each of the phone-based and grapheme-based phoneme set. In these methods, methods with dependent tone always have larger number of basic phonetic units than the methods with independent tone or without using tone (Tab. 5.5). The main advantage of this strategy is that vowel *M* is presented in the form of a group of phonemes. This makes the number of phonetic units

of type vowel much larger than the one presented in the previous strategy and tone attached to this vowel also tends to carry larger pitch information. Because vowel *M* can be constructed from several phonemes, the relation between phonemes in a syllable is somehow modeled which can improve the performance of context-independent ASR of Vietnamese. For this strategy, context-dependent ASR systems are complex to build because the number of phonetic units in each syllable is small and the data needed to cover, say all possible bi-phone or tri-phone based ASR system is large. Fig.5.2 shows the tone and pitch information when analyzing syllable 'TOÁN' using five methods of the vowel-based strategy.

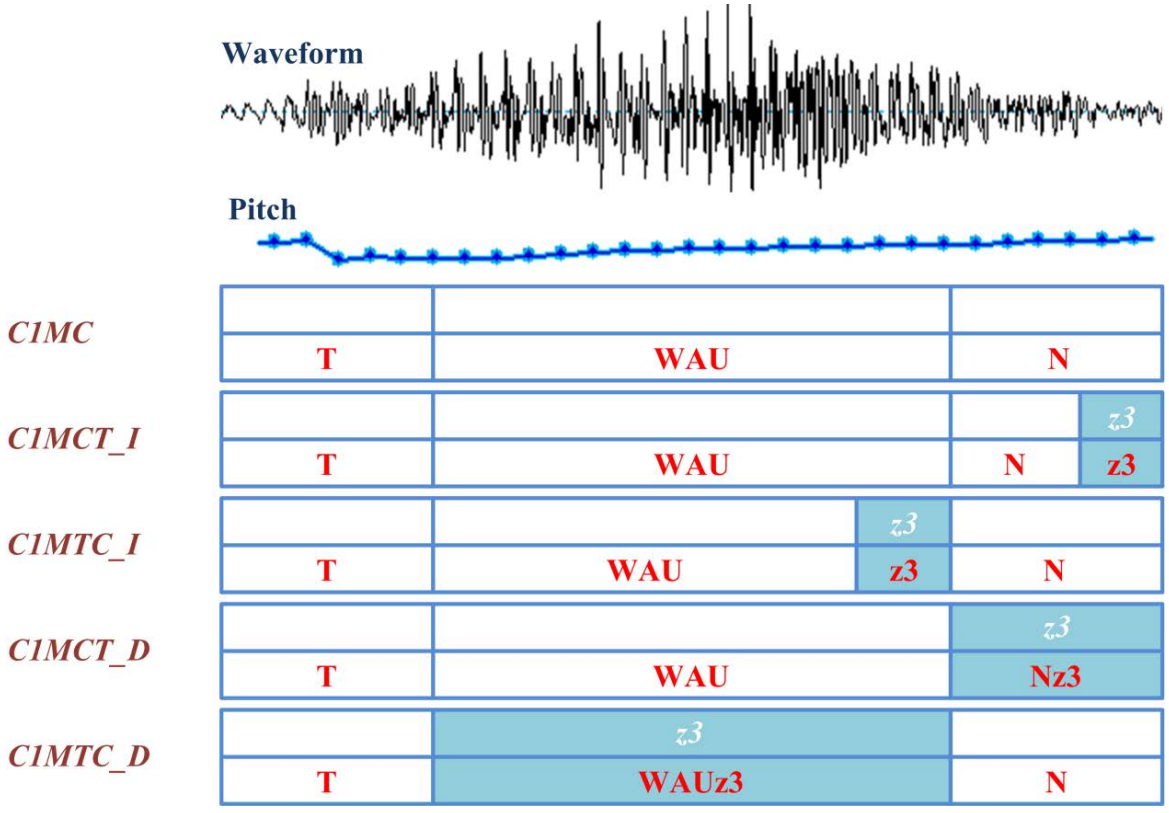


Fig.5.2: Analysis of syllable 'TOÁN' in vowel-based strategy.

5.3.2.1 Without tone analysis: C1MC

For this method, a syllable can be analyzed into at most three basic phonetic units in which the vowel *M* can be monophthong, diphthong or triphthong. For syllables of form *[C1]V[C]* (see sections 3.2.4 and 5.2), the medial *w* and semivowel phoneme of *C2* are missing, and so the vowel *M* will be main vowel *V* which can be one of the monophthongs or diphthongs presented in Tab. 5.2. For syllables of forms *[C1]wV* or *[C1]V[C2]*, the

vowel *M* can be the combination of medial *w* with main vowel *V* or main vowel *V* with semivowel phoneme of *C2* which is one of the diphthongs presented in Tab. 3.4 or triphthongs presented in Tab. 3.5. For syllables of form *[C1]wVC2*, the vowel *M* is constructed from three phonemes *w*, *V* and semivowel of *C2* which is presented as a triphthong shown in Tab. 3.5. The complete sets of basic phonetic units of this method are shown in Tab. A.11 and Tab. A.12. Comparing to method *C1wVC2* of the phoneme-based strategy, this method has larger number of basic phonetic unit because of the presence of not only monophthong but also diphthong and triphthong. The advantage of this method is that the phonetic unit (vowel *M*) contains some kind of context dependent property in the form of biphone or triphone (combination of two or three vowel), and so, more information incident to syllable can be modeled.

Tab. 5.9 shows an example of analyzing the Vietnamese syllable 'TOAN' into its corresponding phonetic units. As in the case of method *C1wVC2*, there is only one pronunciation dictionary entry for different tonal syllables with the same base syllable. Fig.5.2 also shows that vowel 'WAU' carries larger information (pitch, waveform, etc.) in comparison with main vowel 'AU' in method *C1wVC2*.

Tab. 5.9: Analyzing of syllable 'TOAN' using C1MC method.

TOAN		**Phone-based pronunciation dictionary entry**
Tone	*Syllable*	*C1R*
z1	TOAN	T WAU N
z2	TOÀN	
z3	TOÁN	
z4	TOẠN	
z5	TOẢN	
z6	TOÃN	

5.3.2.2 Independent tone analysis: C1MCT_I, C1MTC_I

For this analyzing approach, tone is considered as an independent phonetic unit with the same role as other phonetic units in a syllable. The two methods of this group (*C1MCT_I* and *C1MTC_I*) contain the same set of basic phonetic unit (see Tab. A.13 and Tab. A.14) in which the only difference is the position of tone in syllable. In these methods, a syllable can be analyzed into at most four basic phonetic units where tone is put after the vowel *M* or after the final consonant *C* and is the only tonal phonetic unit of the syllable; other phonetic units including *C1*, *M* and *C* do not carry any tone information. As in the case of *C1wVC2T_I*, *C1wVTC2_I* and *C1wVTC2T_I* methods, there are only six tonal phonetic

units (‘z1’, ‘z2’, ‘z3’, ‘z4’, ‘z5’, ‘z6’) for each of the method in this analyzing approach. This makes the basic phonetic unit sets of these methods are nearly the same as the method *C1MC* with a difference is only from the addition of six tonal phonetic units. Tab. 5.10 and Fig.5.2 show the differences between these two methods and other methods of phoneme-based strategy.

Tab. 5.10: Analyzing of syllable ‘TOAN’ using C1MCT_I and C1MTC_I methods.

TOAN		**Phone-based pronunciation dictionary entry**	
Tone	*Syllable*	*C1MCT_I*	*C1MTC_I*
z1	TOAN	T WAU N z1	T WAU z1 N
z2	TOÀN	T WAU N z2	T WAU z2 N
z3	TOÁN	T WAU N z3	T WAU z3 N
z4	TOẠN	T WAU N z4	T WAU z4 N
z5	TOẢN	T WAU N z5	T WAU z5 N
z6	TOÃN	T WAU N z6	T WAU z6 N

5.3.2.3 Dependent tone analysis: C1MCT_D, C1MTC_D

Two methods *C1MCT_D* and *C1MTC_D* with their corresponding phonetic unit sets shown in Tab. A.15 to Tab. A.18 are grouped into this category. For these methods, a syllable can be analyzed into at most three phonetic units in which tone is attached to the vowel *M* or the final consonant *C* to form a tonal phonetic unit and other phonetic units do not carry any tone information. It is easy to see that these methods integrate tone into syllables the same way as in *C1wVC2T_D* and *C1wVTC2_D* methods but on larger number of basic phonetic units. This makes the number of total basic phonetic unit of these methods is always larger than other methods in the same strategy or in the phoneme-based strategy. Tab. 5.11 and Fig.5.2 show the differences between these two methods and other methods described above.

Tab. 5.11: Analyze syllable ‘TOAN’ using C1MCT_D and C1MTC_D methods.

TOAN		**Phone-based pronunciation dictionary entry**	
Tone	*Syllable*	*C1MCT_D*	*C1MTC_D*
z1	TOAN	T WAU Nz1	T WAUz1 N
z2	TOÀN	T WAU Nz2	T WAUz2 N
z3	TOÁN	T WAU Nz3	T WAUz3 N
z4	TOẠN	T WAU Nz4	T WAUz4 N
z5	TOẢN	T WAU Nz5	T WAUz5 N
z6	TOÃN	T WAU Nz6	T WAUz6 N

5.3.3 Rhyme-based strategy

For this strategy, the basic phonetic units to build acoustic models are initial consonant *C1* and rhyme *R*. Works on Chinese language showed that this strategy can obtain good results for syllable-based language, and so, it can be applied on ASR of Vietnamese. This strategy consists of three basic methods for each of the phone-based and grapheme-based phoneme set. The main advantage of this strategy is that rhyme *R* is presented in form of the largest group of phonemes in comparison with other strategies. This makes the number of phonetic units of this strategy much larger than the one presented in the previous strategies and tone attached to the rhyme *R* is supposed to carry the complete pitch information. Again, methods with dependent tone always have larger number of basic phonetic units than the methods with independent tone or without using tone (Tab. 5.5). The difficulties when applying this strategy is that it needs a larger data to cover all possible basic phonetic units and a good method to model the relation between components of syllables. Fig.5.3 shows the tone and pitch information when analyzing syllable 'TOÁN' using three methods of the rhyme-based strategy.

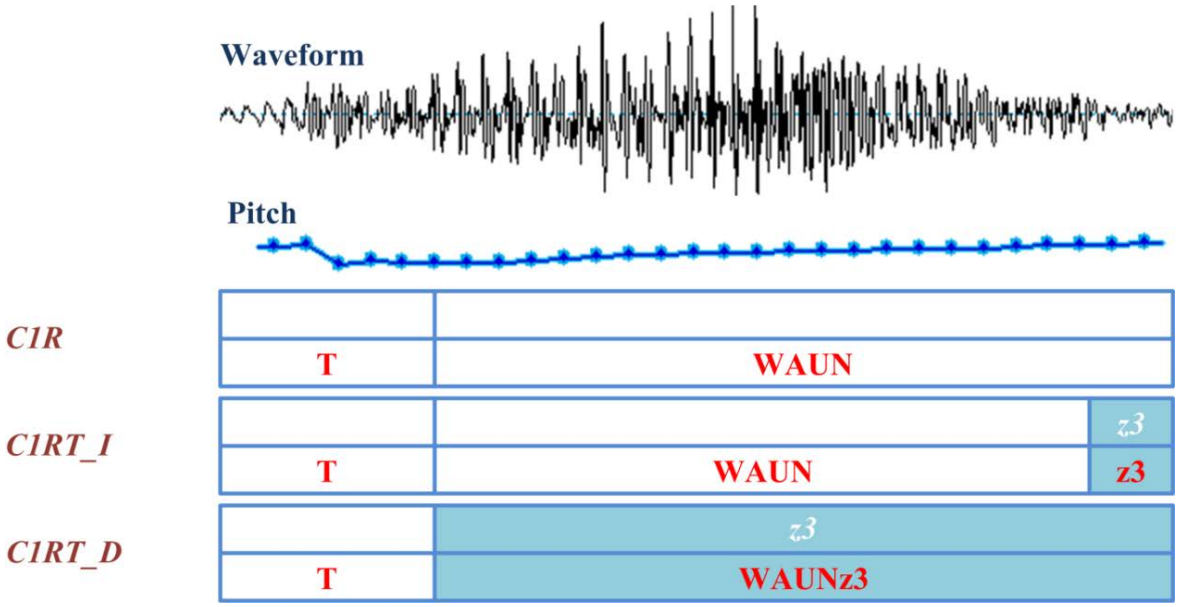


Fig.5.3: Analysis of syllable 'TOÁN' in rhyme-based strategy.

5.3.3.1 Without tone analysis: C1R

For this method, a syllable can be analyzed into at most two basic phonetic units in which the rhyme *R = [w]V[C2]* can be mono-phone, di-phone, tri-phone or even tetra-phone.

Note that medial *w* is always mono-phone, main vowel *V* can be mono-phone or di-phone, and final phoneme *C2* is always mono-phone. Tab. A.19 and Tab. A.20 list the total basic phonetic unit for this analyzing method. Comparing to *C1wVC2* and *C1MC* methods of the previous strategy, this method has larger number of basic phonetic unit because of the presence of multiple combinations of phonemes to form rhyme *R*. For context-independent ASR systems, it is definitely an advantage because the rhyme contains some kind of context dependent property which can be modeled as a phonetic unit.

Tab. 5.12 shows an example of analyzing the Vietnamese syllable 'TOAN' into its corresponding phonetic units. Fig.5.3 also shows that rhyme 'WAUN' carries larger information (pitch, waveform, etc.) in comparison with main vowel 'AU' for method *C1wVC2* (Fig.5.1) or vowel 'WAU' for method *C2MC* (Fig.5.2).

Tab. 5.12: Analyzing of syllable 'TOAN' using C1R method.

TOAN		**Phone-based pronunciation dictionary entry**
Tone	*Syllable*	*C1R*
z1	TOAN	T WAUN
z2	TOÀN	
z3	TOÁN	
z4	TOẠN	
z5	TOẢN	
z6	TOÃN	

5.3.3.2 Independent tone analysis: C1RT_I

For this analyzing approach, tone is considered as an independent phonetic unit and a syllable can be analyzed into at most three basic phonetic units where tone is put after the rhyme and is the only tonal phonetic unit of a syllable. The phonetic unit set of this method is nearly the same as the method *C1R* with the addition of six tonal phonetic units 'z1', 'z2', 'z3', 'z4', 'z5', 'z6' (see Tab. A.21 and Tab. A.22). Although the phonetic unit sets of *C1R* and *C1RT_I* methods are nearly the same, the properties of the phonetic units are different. Fig.5.3 shows that for the same phonetic unit, method *C1R* uses more information (larger boundaries) than method *C1RT_I* to model phonetic units. It is because tone has occupied a part of this information for the acoustic model training in method *C1RT_I*. Tab. 5.13 shows an example of analyzing the syllable 'TOAN' into its corresponding phonetic units with six different tones.

Tab. 5.13: Analyze syllable 'TOAN' using C1RT_I method.

TOAN		**Phone-based pronunciation dictionary entry**
Tone	*Syllable*	*C1RT_I*
z1	TOAN	T WAUN z1
z2	TOÀN	T WAUN z2
z3	TOÁN	T WAUN z3
z4	TOẠN	T WAUN z4
z5	TOẢN	T WAUN z5
z6	TOÃN	T WAUN z6

5.3.3.3 Dependent tone analysis: C1RT_D

For this method, a syllable can be analyzed into at most two phonetic units in which tone is attached to the rhyme *R* to form a tonal phonetic unit of a syllable, and so the phonetic unit tends to carry the largest information in comparison with the previous methods but the number of phonetic units is large and a larger training data will be needed. Tab. A.23 and Tab. A.24 show the total basic phonetic units of this method. Tone information incident to the rhyme *R* of this method is supposed to be fully described. Tab. 5.14 and Fig.5.3 show an example of analyzing the syllable 'TOAN' into its corresponding phonetic units.

Tab. 5.14: Analyzing of syllable 'TOAN' using C1RT_D method.

TOAN		**Phone-based pronunciation dictionary entry**
Tone	*Syllable*	*C1RT_D*
z1	TOAN	T WAUNz1
z2	TOÀN	T WAUNz2
z3	TOÁN	T WAUNz3
z4	TOẠN	T WAUNz4
z5	TOẢN	T WAUNz5
z6	TOÃN	T WAUNz6

5.3.4 Syllable-based strategy

In this strategy, the syllables are not analyzed into smaller components and so it uses all the information available to model each syllable. Because the total number of popular Vietnamese syllable is about 7000 to 8000, it is impractical to construct LVCSR system using this strategy. Isolated word or isolated syllable speech recognition is more suitable for this strategy. The three methods presented in this strategy consist of the whole syllable without tone (S) and the whole syllable with tone (ST_D, ST_I). Fig.5.4 shows the visual information of these methods.

In this work, this strategy is used to evaluate the task of audio only and audio-visual isolated word speech recognition.

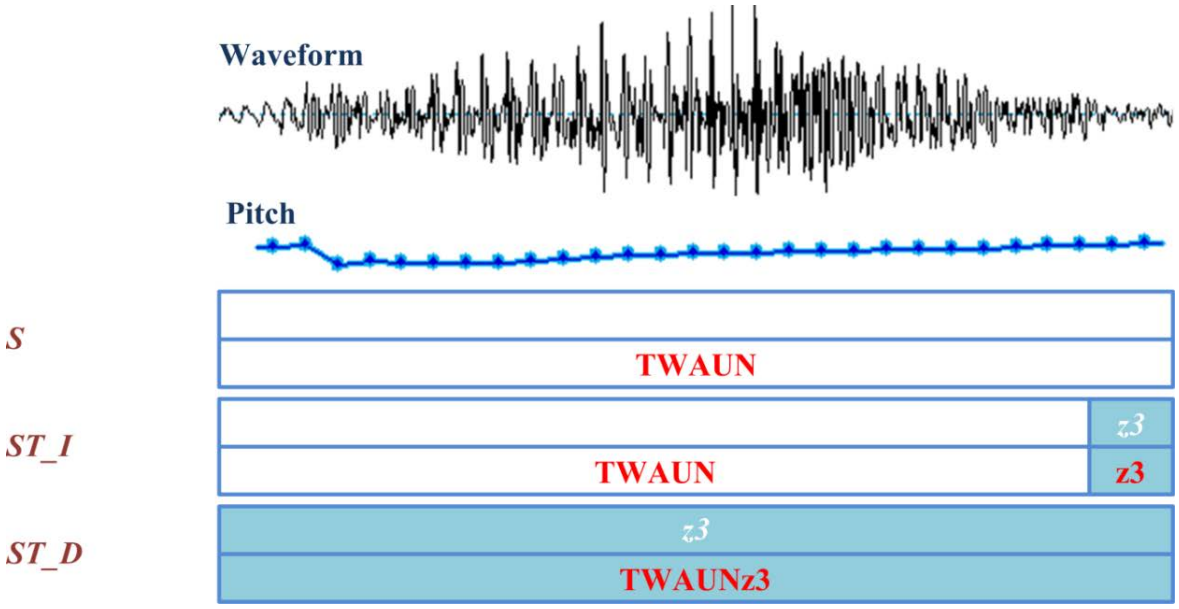


Fig.5.4: Analysis of syllable 'TOÁN' in syllable-based strategy.

Chapter 6

Text, Audio and Audio-Visual Databases

6.1 Building of Vietnamese text corpus from the Internet

In this section, a method for collecting a large, general purpose text corpus of Vietnamese by exploiting the source of linguistic data from the Internet [P2] is described. There are two basic steps to obtain this corpus. First, a small text corpus will be collected from a well-known Internet's resource: the Wikipedia. Because the corpus resulted from this resource is not large and diverse enough, this corpus is used to extract seed words only. From these seed words, web queries are generated and fed to the search engine to collect only links available for Vietnamese language. After downloading and extracting all the useful text from these links, the final text corpus that good enough for multipurpose research is obtained. Another specific purpose text corpus is also created by extract text from websites which are rich of literature and news resource.

6.1.1 Vietnamese text corpus from Wikipedia

To extract a Vietnamese text corpus from Wikipedia, a Wiki data dump of Vietnamese is needed. A Wiki data dump is a single large file containing all the articles on the Wikipedia website. From the Wikipedia's service, there is a list of data dumps of Vietnamese can be downloaded, but only the dump for page's articles is used in this work because it contains all the articles in which useful text can be extracted. The plain text is then collected from the dump of size 670 MB using a modified version of the 'Wikipedia2Text' tool. Because there are many articles do not contain any useful text, all of these articles have to be filtered out. One possible solution is to keep only articles which have more than 500 syllables or xml files with size larger than 10 KB. Although there will be articles meet the above

conditions but still do not have any useful text, their effect on the statistic of syllables in the corpus is not significant. Some statistics of this corpus are shown in Tab. 6.1.

Tab. 6.1: Statistics of Wikipedia text corpus.

Size of Wiki XML dump (Mb)	670
Size of the resulted text corpus (Mb)	191
Number of Vietnamese syllables	1,274,662

6.1.2 Extracting of general purpose text corpus

To build a general purpose text corpus from the Internet, a frequency list of words in the Wikipedia text corpus is first created. Because Vietnamese words are formed from syllables, which are separated by space, it is difficult to specify the boundary of Vietnamese words. To overcome this obstacle, a lexicon of Vietnamese with 73901 common words is used to build the frequency list of words. The frequency of occurrence of all the words in this lexicon will be computed from the Wikipedia text corpus resulted in a frequency list of all common Vietnamese words which are sorted in decrease order for the next processing step. Using the resulted frequency list, a set of 5000 seed words is then selected of which the following requirements have to be met:

- The frequency of occurrence of the seed words has to be not too high because the high frequency seed words tend to present in most of the Web pages extracted from the Internet when these words are fed to the search engine. And so, the possibility of page duplication is high and the number of useful collected links will be small. On the other hand, many letters in the Vietnamese writing system use Latin symbols which are the same to those in writing system of many other Western countries. This fact leads to a problem that the very high frequency seed words, which are usually short, may be confused with words from other languages. E.g. Vietnamese word “do” (because) may be confused with English word “do”, Vietnamese word “bay” (to fly) may be confused with English word “bay”, etc.
- These seed words should be sufficiently general because this work tends to create a large multipurpose text corpus. Seed words, which are too specific, may result in small number of links returned from the search engine, and so the final text corpus will be small and not diverse enough.

- Seed words have to contain at least one of the specific Vietnamese symbols: ‘ă’, ‘â’, ‘đ’, ‘ê’, ‘ô’, ‘ơ’, ‘ư’ and/or the last *five diacritic symbols* that mark the tone in the accent set. Some examples of seed words of this type are: ‘vở’, ‘tá’, ‘tý’. As a result, seed words with those symbols are specific for Vietnamese word, and so, the web pages resulted from the search engine when dealing with these seed words are most likely containing only Vietnamese text.

To select seed words satisfying the above requirements, the first 1000 words with very high frequency of occurrence in the frequency list are disposed of. Then the first 5000 words of the remaining words that meet the third requirement are selected (word with the highest frequency was chosen first). Tab. 6.2 shows a list of the first 20 seed words with their corresponding frequency of occurrence. Note that both monosyllabic and polysyllabic words are presented in this list.

Tab. 6.2: Example of seed words.

Seed word	Frequency	Seed word	Frequency
ổ bi	5877	nổ	5875
bóng đá	5873	ấy	5865
sư đoàn	5864	phận	5843
đem	5843	phù	5832
đầy	5823	tương tự	5820
ánh	5817	đen	5815
chuyến	5805	tội	5803
mạc	5800	thay thế	5793
suốt	5788	hổ biến	5783
bảng	5783	phổ biến	5781

The BootCat’s query generation tool is then used to generate web queries of seed words obtained from the previous step. Each query is formed from *N* seed words and met two requirements: there are not any two queries that are identical or a permutation of each other and *N* seed words are randomly selected without replacement. The problem of this step is that how to determine the length of seed words *N*. If the number of seed words for a query is large, the number of queries that can be formed from these seed words is large and the probability that Web links return from the search engine given these queries contain only Vietnamese text is high. However, when the length *N* of each query is large, the number of Web links returned from the search engine is small. As a result, the requirement is to select a query length *N* that can balance those effects. One possible solution is to select the largest query length *N* for which the hit counts (links returned from a search engine) for

most queries (say, 90%) is more than ten. The following algorithm is used to determine the best query length for Vietnamese.

Algorithm 6.1: Compute query length

1. Set $N = 1$, number of hit per query equal 10
2. Generate 100 queries using N seeds words per query
3. Sort queries by the number of hits they get
4. Count the number of hits H for the first 90 queries
5. If H < 900 return N -1
6. $N = N + 1$, go to step 2

Using this algorithm, $N = 2$ is the best query length returned from the search process. Tab. 6.3 shows that when query length larger than 2, the total hits of 90 best queries are smaller than 900 and reduce when larger query length is used. Once the query length was established, around 35000 queries are generated using 5000 seed words obtained from the previous step. These queries are fed to the Yahoo's API search engine to get the first ten hits for each query. The Web links returned from the search engine are checked for duplication resulted in a total of 66838 links. Then links with size from 5 KB to 2 MB are downloaded for the next processing step. By extracting all the useful text from the downloaded Web pages, a text corpus that is large and diverse enough for multipurpose research is obtained. Some statistics of this raw corpus are show in Tab. 6.4.

Tab. 6.3: Statistics of query length selection.

Query length N	1	2	3	4	5
Total hits	1,000	964	879	749	630
Total queries with 10 hits per query	100	94	85	72	61
Total hits of 90 best queries	900	900	875	749	630

Tab. 6.4: Statistics of raw data.

Number of unique URLs Collected	66,838
Number of URLs after Filtering	53,009
Size of the resulted text corpus (Mb)	277
Number of sentences	2,084,088
Number of Vietnamese syllables	53,943,274

6.1.3 Extracting of specific purpose text corpus

In the previous section, a text corpus for general purpose was collected using the Internet resources. For this corpus, the number of Web page returned from the search engine is large

and belongs to varying categories such as book, newspaper, law, cultural, history, etc., and so the extracted text corpus may not good enough for some specific tasks such as news dictating system. The second problem of this corpus is that it relatively small. The size of raw text without filtering foreign or strange words, duplicate sentence, etc., is 277 Mb with 2084088 sentences and 53943274 Vietnamese syllables. For all of the above reasons, another text corpus is collected for Vietnamese LVCSR task. This text corpus is tended for specific purpose with text mainly in the field of news and literature, and so several Vietnamese websites which are rich of newspaper and book resource are collected. The list of all websites is shown in Tab. 6.5. For each website, only Web pages with size from 5 KB to 2 MB are downloaded and extracted all useful text. The resulted text corpus has raw text with size of 955.4 Mb.

Tab. 6.5: Website for collecting text corpus.

vietnamnet.vn	vnexpress.net	nhanhdan.com.vn
dantri.com.vn	vnthuquan.net	vanhoc.xitrum.net

6.1.4 Filtering of text corpora

The two text corpora collected from the previous sections contain only raw text extracted from web pages and need to be filtered out unwanted text such as duplicate sentence, foreign word, abbreviation, number and other strange scripts incident to the raw text. The procedure below is used to filter raw text of the text corpora.

Procedure 6.1: Filtering of raw text

1. The raw text is segmented into sentences and sentence will be the basic unit when filtering unwanted text.
2. For all sentences in the raw text, only sentences with 100% Vietnamese syllable are selected. Sentences which contain one or more foreign word, number, abbreviation, strange script, etc., are filtered out for the next processing step. For the selected sentences, all duplicate sentences will be removed and the remaining sentences are stored in set 1 (*VN only* set) of the processed corpus.
3. The sentences which are filtered out from the previous step will be further processed to collect useful text. In this step, all sentences which contain only 20% of foreign word, number, abbreviation, strange script, etc., are selected and the rest will be

removed. All duplicate sentences will also be removed and the remaining sentences are stored in set 2 (*VN mix* set) of the processed corpus.

The resulted text corpora are used to build language models for the continuous speech recognition tasks. Tab. 6.6 to Tab. 6.8 show some statistics of the text corpora of three various categories (general purpose, news and literature). Each category contains text of two groups: the first group consists of sentences with Vietnamese syllables only and the second group consists of sentences with both Vietnamese syllables and other tokens such as digit, foreign word, abbreviation, etc. Tab. 6.9 shows the statistics of the total text corpora which are the combination of text corpora of the above three categories.

Tab. 6.6: Statistics of the filtered general purpose text corpus.

	VN only	*VN mix*
Raw text size (Mb)	277	
Filtered text size (Mb)	71.3	185
No. of sentence	664,942	1,163,802
No. of syllable	12,443,710	30,782,462
No. of foreign word	0	1,757,193

Tab. 6.7: Statistics of the filtered specific text corpus (news).

	VN only	*VN mix*
Raw text size (Mb)	570.5	
Filtered text size (Mb)	75.5	356
No. of sentence	561,126	2,174,656
No. of syllable	13,122,518	58,124,913
No. of foreign word	0	4,629,387

Tab. 6.8: Statistics of the filtered specific text corpus (literature).

	VN only	*VN mix*
Raw text size (Mb)	384.9	
Filtered text size (Mb)	263	117
No. of sentence	2,573,061	996,252
No. of syllable	46,170,702	19,826,804
No. of foreign word	0	906,483

Tab. 6.9: Statistics of the total filtered text corpora.

	VN only	*VN mix*	*VN total*
Raw text size (Mb)	1,232.4		
Filtered text size (Mb)	409.8	658	1,067.8
No. of sentence	3,799,129	4,334,710	8,133,839
No. of syllable	71,736,930	108,734,179	180,471,109
No. of foreign word	0	7,293,063	7,293,063

6.2 Collecting of speech corpus for LVCSR task

In developing of a speech recognition system, the requirement of an available speech corpus is crucial. For some well-known languages such as English, Spanish, etc., there are already many speech corpora that can be used either for research purposes or commercial purposes. On the other hand, for some under-resourced languages, the availability of a standard and reliable speech corpus is one of the biggest obstacles when developing speech recognition systems. In Vietnam, the research on speech recognition began about twenty years ago. However, because of the limitation of knowledge, budget and time, the available corpora are small and not standard. To solve this problem, a speech corpus for LVCSR of Vietnamese is collected from the Internet resource. Initial work of speech corpus selection is shown in [P2]. First, sound files from some main websites which are rich of speech data are downloaded and converted into required format. Then only good utterances in those speech data are selected and manually transcribed to obtain a total of 24871 utterances with the length of 50 hours 22 minutes. The number of speaker in this speech corpus is 196 (69 male speakers and 127 female speakers).

This corpus contains speech mainly of type: story reading, news report, weather forecast, conversation and has the following properties: it covers three main dialects of Vietnamese language (north, south and center), has many range of speaking rate (syllable per minute) and contains varying type of background noise in which each speech data file has each own noise condition (room, studio, outdoor, etc.). With the diversity of this speech corpus, the effect of tone can be examined totally in the experiments on LVCSR of Vietnamese.

6.3 Designing of audio-visual speech corpus

To select sentences for recording a Vietnamese audio-visual speech corpus, an original sentence set is needed. This sentence set has to be selected to meet some requirements: each sentence in the original sentence set has to contain at least 3 and at most 15 syllables, and these sentences do not contain any foreign words, numbers and abbreviations. To create an original sentence set, the general purpose text corpus collected from the previous section is used. From this text corpus, only sentences that meet the above requirements are chosen

resulted in a set of 540744 sentences with 43356 distinct words as an original sentence set (Tab. 6.10).

Tab. 6.10: Original sentence set statistics.

Number of sentences	540,744
Length of sentences	3 to 15
Number of syllables in sentence set	4,779,602
Number of distinct words	43,356

From the original sentence set, two selecting blocks are applied to select the sentences for audio-visual speech recording. The first block is used to produce the smallest sentence set that covers all units in the selected phonetic units of the original sentence set. The second block is used to select phonetically balanced sentences based on the frequency of occurrence of phonetic units in the original sentences. These two blocks can be applied independently to produce a sentence set of its own or can be combined together to create a better sentence set which contains all selected phonetic units in the original sentences and is phonetically balanced.

The selection procedure of the first block is shown in Fig.6.1(a). The original sentence set and selected phonetic units are passed through this block to produce the desired sentence set. This block can also use covered phonetic units resulted from the second block as an extra input. The first block will successively select sentences as follows: at any stage, the sentence with the largest distinct phonetic unit count of the remaining uncovered units is selected and all distinct phonetic units in that sentence are covered. If two or more sentences have the same count, the sentence with a higher phonetically balance contribution to the selected sentences will be selected. The selection procedure is repeated until all phonetic units are covered.

The second block is described in Fig.6.1(b). The input to this block is the original sentence set and the selected sentences resulted from the previous step. Like the first block, a selection procedure is applied to produce the phonetically balanced sentence set. At any stage, a score is calculated for each sentence to present the contribution of phonetic units in that sentence in the original sentence set. The sentence with the best score will be selected and moved to the current selected sentences. Repeat this procedure until the number of desired sentences is selected. The score is defined differently to the score used in the first block. The purpose is to ensure that the selected sentences contain phonetic units

correspond to their frequency of occurrence in the original sentence set. To calculate the score for each sentence, the score for each distinct phonetic unit in that sentence is calculated and sum over for all distinct phonetic units. The score for each distinct phonetic unit in the examined sentence was computed by taking the absolute value of the different between relative frequencies of occurrence of that phonetic unit in the original sentence set and in both the selected and examined sentences. The sentence with the highest score must be selected.

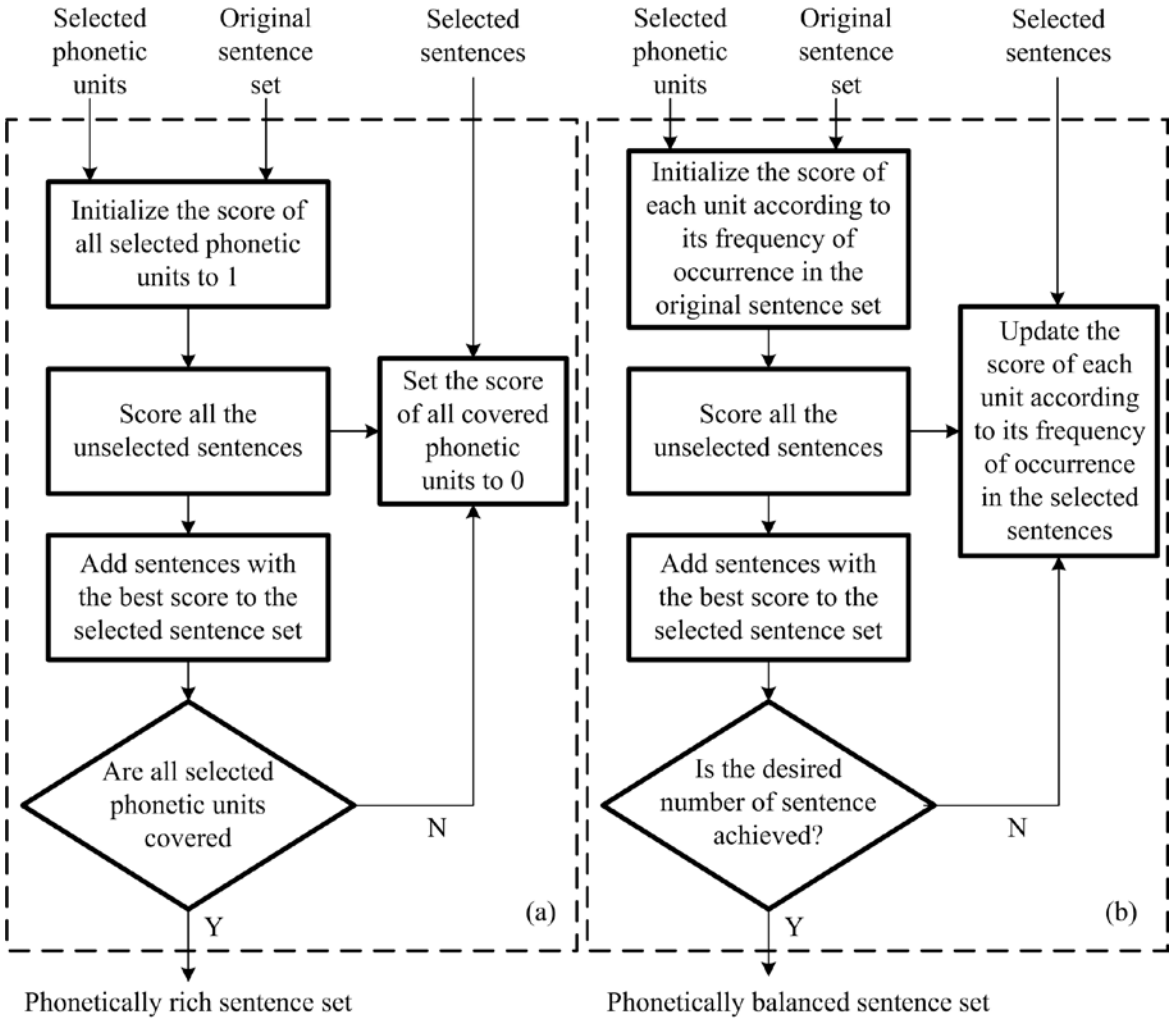


Fig.6.1: Block diagrams of sentence selection procedures: a) Sentence selection block 1; b) Sentence selection block 2.

For continuous speech recognition experiments, the sentence selection block 2 is first used to select 50 adaptation sentences in which there are no duplicated sentences. Then two sentence selection blocks are combined to select another 2500 sentences that meet two requirements: they have to cover all selected phoneme types in the original sentence set and

are phonetically balanced. The audio-visual speech corpus recorded using the selected sentences can be used for developing and evaluating speaker-independent speech recognition system and also examining the recognition of a specific speaker for Vietnamese. The above sentences are selected using monophone as phonetic unit. In [P4] the same procedure was applied to select sentences, but in this case triphone was used as phonetic unit.

In the audio-visual speech corpus, an isolated word speech data is also recorded for applications such as PC commands, device control in SmartRoom, etc., so, the isolated words are selected to meet the above requirements. The list of all isolated words is shown in Tab. 6.11.

Tab. 6.11 : 50 isolated words for audio-visual speech data recording.

Digit	**Noun**	**Preposition**	**Verb**	
không (0)	đèn (light,lamp)	trái (left)	kiểm tra (check)	lập lại (repeat)
một(1)	quạt máy (fan)	phải (right)	xác nhận (verify)	nhận dạng (recognize)
hai (2)	máy tính (computer)	trên (above)	đồng ý (yes)	hủy (cancel)
ba (3)	máy in (printer)	dưới (below)	bắt đầu (start)	chọn (select)
bốn (4)	ti-vi (television)	trước (front)	dừng (stop)	nghe (hear)
năm (5)	cửa (door)	sau (behind)	hoạt động (run, function)	thu (record)
sáu (6)	cửa sổ (window)	giữa (middle)	thực hiện (do, perform)	lưu (save)
bảy (7)	số (number)		mở (open, turn on)	
tám (8)	tên (name)		đóng (close)	
chín (9)	lệnh (command)		tắt (turn off)	
	lựa chọn (option)		tăng (up, increase)	
	âm lượng (volume)		giảm (down, decrease)	
	tốc độ (speed)		tiếp tục (continue)	

The audio-visual speech data contains frontal face of 50 speakers. Each speaker is asked to utter the same 50 isolated words, 50 specific sentences and 50 general sentences in front of a camera in relatively clean condition (quiet room with some type of background noises like computer fan, etc.) resulted in a total 2500 isolated word utterances, 2500 specific sentence and 2500 general sentence utterances. The video is sampled at a rate of 30 frames per second (fps) and the resolution of each video frame is 640 x480 pixels, 24 bits per pixel. The audio is sampled at a rate of 11025 Hz, 8 bits per sample.

Chapter 7

Audio Speech Recognition of Vietnamese

7.1 Building language model for LVCSR of Vietnamese

For speech recognition tasks, there are two types of LM: grammar and statistical LM. The grammar LM is very simple and more suitable for applications such as command and control. Because this research tends to deal with LVCSR of Vietnamese, statistical *n*-gram LM will be examined.

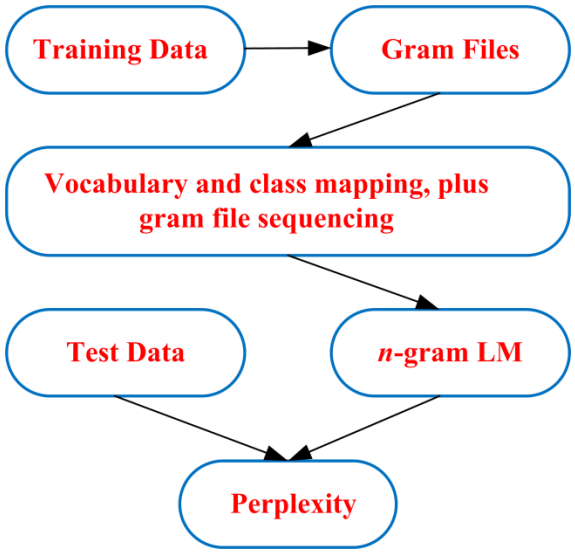


Fig.7.1: Constructing and testing *n*-gram LM.

As can be seen in Fig.7.1, a *n*-gram LM is constructed as follows: Firstly, the *n*-grams (a sequence of *n* symbols) are counted and stored in *gram* files. And then some words may be mapped to an out of vocabulary class or other class mapping may be applied. Finally, the gram files are used to compute *n*-gram probabilities which are stored in the LM file.

The performance of a LM applied on unseen test data can be evaluated by computing a measure called *perplexity*.

In practice, it is difficult to build a single LM which is robust to varying applications. For example, an LM built on conversation text would be a good predictor for phone answering system but the same LM would be a poor predictor for dictating news reports or literature. For this work, LMs built from various types of text corpus are examined.

When building LM for LVCSR of Vietnamese tasks, two main problems have to be solved including LM type and system's vocabulary size. There are two types of LM that can be used for ASR of Vietnamese: word-based LM and syllable-based LM. The construction of word-based LM is more difficult and depended on the existence of a method for segmenting Vietnamese word in a given sentence or paragraph. The vocabulary size of this LM type can also be large (up to 70000 words). On the other hand, the construction of syllable-based LM is much simple and the vocabulary size is also small. Although the number of pronounceable Vietnamese syllables is about 20000, the most popular syllables that are used in the writing and speaking system of Vietnamese are only from 7000 to 8000.

7.1.1 Syllable-based LM construction

For simplicity, this work is more concentrated on the effect of syllable-based LM because the training of this LM type is straightforward for Vietnamese. As mentioned in the previous sections, syllables are written separately by space in the Vietnamese writing system, and so the task of syllable segmentation is easy to accomplish. Also knowing that, in modern Vietnamese, there are about 7000 to 8000 syllables which are used frequently in writing as well as in speaking. All of the LMs will be trained using the text corpora described in section 6.1 in which various factors that affect the perplexity of LM such as vocabulary size, text corpus category, LM smoothing method, etc., will be examined.

First, to evaluate the effect of different text corpus categories, all the text corpora presented in section 6.1.4 are utilized. The constructed LMs will have the vocabularies of size 6000 and 7000 syllables. These syllables are selected based on their frequency of occurrence in the text corpus where syllables with the highest count will be selected and stored in the system's dictionary. Also, another vocabulary that contains all 5741 distinct

syllables occurring in the training part of the speech corpus's transcription is used to build LM. All the LMs are trained using Good-Turning smoothing method. The testing data contains all 24871 sentences in the LVCSR speech corpus (section 6.2). These sentences consist of a total of 581221 syllables which presented by 5865 distinct syllables.

The perplexities of the constructed LMs are shown in Tab. 7.1 to Tab. 7.4. Because the testing data is mainly from the domain of literature, the LMs trained from the text corpora of literature category give better results than the LMs trained from the other two categories (general purpose and news). The results also show that using text corpora with all sentences containing the Vietnamese syllable only (*VN only* set) gives better perplexities than the case where both Vietnamese syllable and unknown token are presented in the text corpora (*VN mix* set). One exception is the case of bi-gram LM constructed using text in literature domain but the difference in perplexity is not too much. Note that, although the vocabularies contain only 6000 to 7000 syllables, they cover more than 90 percent of the training text corpora. It is also interesting to see that LMs based on statistic to select the syllables perform better than the case where all syllables in the transcription of training speech corpus are selected.

Tab. 7.1: LM test on general purpose text corpus.

Vocabulary size	**Perplexity VN only**		**Perplexity VN mix**	
	bi-gram	*tri-gram*	*bi-gram*	*tri-gram*
6,000	282.695	223.174	333.048	255.299
7,000	286.079	226.135	336.974	258.698
5,741	289.266	228.958	340.970	262.182

Tab. 7.2: LM test on specific text corpus (literature).

Vocabulary size	**Perplexity VN only**		**Perplexity VN mix**	
	bi-gram	*tri-gram*	*bi-gram*	*tri-gram*
6,000	271.829	190.435	268.552	203.623
7,000	274.561	192.664	271.350	206.018
5,741	277.379	194.978	274.189	208.467

Tab. 7.3: LM test on specific text corpus (news).

Vocabulary size	**Perplexity VN only**		**Perplexity VN mix**	
	bi-gram	*tri-gram*	*bi-gram*	*tri-gram*
6,000	386.847	324.860	468.445	373.348
7,000	392.274	329.856	475.095	379.264
5,741	396.864	334.169	481.417	384.931

Tab. 7.4: LM test on total text corpus.

Vocabulary size	Perplexity VN only		Perplexity VN mix		Perplexity VN total	
	bi-gram	*tri-gram*	*bi-gram*	*tri-gram*	*bi-gram*	*tri-gram*
6,000	246.843	160.330	298.843	199.403	254.966	154.780
7,000	249.457	162.336	302.190	202.025	257.679	156.756
5,741	252.159	164.424	305.772	204.849	260.533	158.855

To further examine the effect of vocabulary size as well as the effect of various smoothing methods on LM constructing, the *VN only* text corpus presented in Tab. 6.9 is used for LM estimation and the same testing data as the previous experiment is used for evaluation. All LMs will be trained using vocabulary of sizes 6000 and 7000 syllables. These syllables are selected based on their frequency of occurrence in the *VN only* text corpus. Another vocabulary that contains all 11017 distinct syllables occurring in the *VN only* text corpus will be examined. As in the previous experiment on LM, the same vocabulary of size 5741 syllables is also studied. In this experiment, the SRILM toolkit [131] will be used to construct all LMs in which three different smoothing methods including Good-Turning, Kneser-Ney and Witten-Bell are taken into consideration. These smoothing methods are based on the backoff models to compute the estimated conditional probability of a word. Besides, for the Kneser-Ney smoothing, these probabilities are also computed using interpolated model. For each smoothing method as well as vocabulary size, both of the bi-gram and tri-gram LMs are trained.

Tab. 7.5 to Tab. 7.8 show the effect of different vocabulary sizes on the perplexity of LMs. For vocabularies where syllables are selected using statistic (LMs with 6000 and 7000 syllables), LMs with smaller vocabulary size achieve better perplexities. This is true both for bi-gram and tri-gram LMs. Also, like the previous experiment, LMs based on vocabulary containing only syllables occurring in the training part of speech corpus (LMs with 5741 syllables) give better results only in the case where all syllables in the training text corpus are used.

The effect of different smoothing methods is then presented in Tab. 7.9 to Tab. 7.12. It is easy to see that Kneser-Ney smoothing using interpolated model gives the best results for all possible vocabulary sizes. For LMs using backoff model, Witten-Bell smoothing provides better perplexities than Good-Turning smoothing and even better than Kneser-Ney smoothing in the case of bi-gram LM. But in the case of tri-gram LMs, Kneser-Ney

smoothing using backoff model or interpolated model all give better perplexities than Good-Turning and Witten-Bell smoothing. The effect of these LMs will be examined in the task of LVCSR of Vietnamese.

Tab. 7.5: LM test using Good-Turning smoothing.

Vocabulary size	Perplexity	
	bi-gram	*tri-gram*
6000	246.040	160.130
7000	248.388	161.688
11017 (All)	249.998	162.738
5741	249.814	162.637

Tab. 7.6: LM test using Witten-Bell smoothing.

Vocabulary size	Perplexity	
	bi-gram	*tri-gram*
6000	244.675	157.851
7000	246.821	159.258
11017 (All)	248.390	160.272
5741	247.942	160.010

Tab. 7.7: LM test using Kneser-Ney smoothing.

Vocabulary size	Perplexity	
	bi-gram	*tri-gram*
6000	245.907	154.401
7000	247.651	155.469
11017 (All)	249.029	156.325
5741	248.528	155.943

Tab. 7.8: LM test using Kneser-Ney smoothing with interpolation.

Vocabulary size	Perplexity	
	bi-gram	*tri-gram*
6000	242.671	141.610
7000	244.624	142.703
11017 (All)	246.197	143.595
5741	245.722	143.319

Tab. 7.9: LM test using vocabulary of 6000 syllables.

Smoothing method	Perplexity	
	bi-gram	*tri-gram*
Good-Turning	246.040	160.130
Kneser-Ney	245.907	154.401
Kneser-Ney interpolation	242.671	141.610
Witten-Bell	244.675	157.851

Tab. 7.10: LM test using vocabulary of 7000 syllables.

Smoothing method	Perplexity	
	bi-gram	*tri-gram*
Good-Turning	248.388	161.688
Kneser-Ney	247.651	155.469
Kneser-Ney interpolation	244.624	142.703
Witten-Bell	246.821	159.258

Tab. 7.11: LM test using vocabulary of all syllables (11017).

Smoothing method	Perplexity	
	bi-gram	*tri-gram*
Good-Turning	249.998	162.738
Kneser-Ney	249.029	156.325
Kneser-Ney interpolation	246.197	143.595
Witten-Bell	248.390	160.272

Tab. 7.12: LM test using vocabulary of 5741.

Smoothing method	Perplexity	
	bi-gram	*tri-gram*
Good-Turning	249.814	162.637
Kneser-Ney	248.528	155.943
Kneser-Ney interpolation	245.722	143.319
Witten-Bell	247.942	160.010

7.1.2 Word-based LM construction

As mentioned in the previous sections, word-based LM is more difficult to construct than syllable-based LM. There are two main factors that affect the performance of this LM type: 1) the selection of multi-syllabic word, and 2) the method to segment multi-syllabic word.

For the first factor, multi-syllabic words can be selected using the standard Vietnamese lexicon or based solely on the statistics of these words in a given text corpus. When using the standard lexicon, the number of words selected can be up to 70000. But in the case of statistical method, the number of possible words selected could be larger. Note that the selected words can be real Vietnamese words (words in the lexicon) or multi-syllabic tokens (group of syllables) which are not presented in the lexicon.

For the second factor, the availability of a stable and accurate word segmented algorithm is the key that determines the performance of word-based LM. There are two main approaches for segmenting multi-syllabic word including linguistic and data-driven approach. In linguistic approach, the knowledge of Vietnamese language will be used to

segment words. It can be done manually by experts or automatically by using some mathematical tool such as conditional random fields (CRFs), support vector machine (SVM), etc. In the case of manual word segmentation, except for rare or difficult multi-syllable words, which can be solved by using only popular words in the Vietnamese lexicon, this task can be accomplished by many Vietnamese researchers and so it is the most stable and accurate method. In the case of using mathematical tool, the available of a training data is needed. The biggest problem of the linguistic approach is that the data labeling process is an unavoidable and time-consuming task. For data-driven approach, the word segmentation process can base solely on the statistical information of words in the text corpus. This approach is affected by many factors including the number of words in vocabulary, the level of word to be segmented (bi-syllable, tri-syllable or more), method to segment word based on statistic, etc.

Because the difficulty of word segmentation task which requires a lot of works to obtain reasonable result, this work uses a simple data-driven approach to segment multi-syllabic token for the tasks of LM construction and LVCSR test. Note that, using this segmenting strategy there will be two types of token presented in the dictionary. The first type contains real Vietnamese words and the second type is multi-syllabic tokens that are not presented in the lexicon. And so, we will call the constructed LMs as multi-syllable-based LMs. For simplicity, only bi-syllabic tokens will be segmented using their frequency of occurrence in the text corpus. Procedure 7.1 shows the segmentation process of a given text corpus.

Procedure 7.1: Bi-syllabic token segmentation

1. Count the frequency of occurrence of all bi-syllabic tokens in the text corpus.
2. Select *N* bi-syllabic tokens with the highest counts and put into the vocabulary.

3. Segmenting a given sentence using tokens in the vocabulary with their corresponding counts as follows: get three neighbor syllables s1, s2, s3 and form two bi-syllabic tokens w1, w2. Get the counts c1, c2 of these tokens in the vocabulary. If $c1 = 0$ and $c2 = 0$, select the next three syllables start at s3. If $c1 = 0$ and $c2 \neq 0$, select the next three syllables start at s2. If $c1 \neq 0$ and $c2 = 0$ or if $c1 \geq c2$, w1 is labeled by putting an underscore between syllable s1 and s2, and select the next three syllables start at s3. If $c1 < c2$, select the next three syllables start at s2. Repeat to the end of sentence.

In this work, the *VN only* text corpus is used for multi-syllable-based LM estimation and the same testing data as the previous experiment is used for evaluation. Both the training and testing data will be segmented using procedure 7.1 in which the number of bi-syllabic tokens *N* will be chosen to have values from 50 to 50000. Tab. 7.13 shows the vocabulary size (total number of monosyllabic and bi-syllabic tokens occurring in the text corpus) and the corresponding actual number of bi-syllabic tokens in the vocabulary. It can be seen that, when the value of *N* is from 5000 to 50000, the actual number of bi-syllabic tokens in the segmented text corpus is smaller than the original bi-syllabic token *N* used to label the corpus. This fact proves that text corpus which is large enough for syllable-based LM may not large enough for multi-syllable-based LM.

For this experiment, the SRILM toolkit is used to construct all bi-gram LMs in which the Kneser-Ney smoothing using interpolated model will be examined. The vocabularies will consist of all tokens occurring in the training text corpus. Tab. 7.13 shows the perplexities of multi-syllable-based LMs estimated using different values of *N*. Note that, the evaluation will be on multi-syllabic level, and not syllabic level, and so the actual performance of these LMs will be totally examined in the task of LVCSR described in section 7.3.3.3.

Tab. 7.13: Multi-syllable-based bi-gram LM test.

N	**Perplexity**	**Vocabulary size**	**Number of actual bi-syllabic token**
50	276.867	11,061	50
80	287.247	11,091	80
100	293.727	11,108	100
500	377.23	11,489	500
1,000	450.526	11,968	1,000
5,000	794.085	15,896	4,999
10,000	1,065.54	20,851	9,995
15,000	1,087.01	26,013	14,992
20,000	1,237.13	30,986	19,965
30,000	1,483.44	40,930	29,909
40,000	1,684.73	50,820	39,800
50,000	1,852.05	60,681	49,661

7.2 Isolated word speech recognition

For experiments on isolated word speech recognition, the audio part of the Vietnamese audio-visual database (Section 6.3) is used in which the training speech contains 50 isolated words uttered by 40 speakers and the testing speech is from the other 10 speakers. In these

experiments, two types of features will be examined: linear prediction filter coefficients (LPC) and mel-frequency cepstral coefficients (MFCC). The feature vectors are sampled at rates of 30 Hz and 100 Hz using two different window sizes 33.3333ms and 25ms respectively. Word-based HMM of all 50 isolated words are trained using feature vectors of 39 dimensions (13 static coefficients and their first and second derivatives). Each HMM has 1 mixture and N states where the number of states N is changed from 3 to 14 to determine the best number of states for the isolated word task. Note that word-based strategy can be considered as the extension of syllable-based strategy (section 5.3.4) where multisyllabic words are trained the same way as monosyllabic words. The HTK toolkit [139] is used both for training and testing purposes.

The results of isolated word speech recognition using all of the above parameters are shown in Fig.7.2 and Tab. 7.14. It can be seen that MFCC coefficient give better result than LPC coefficient and the best number of states used to train HMM for each isolated word is 14. By applying cepstral mean normalization (CMN) on the MFCC coefficients (MFCC_Z), the recognition result is further improved and obtains the highest recognition rate of 97%.

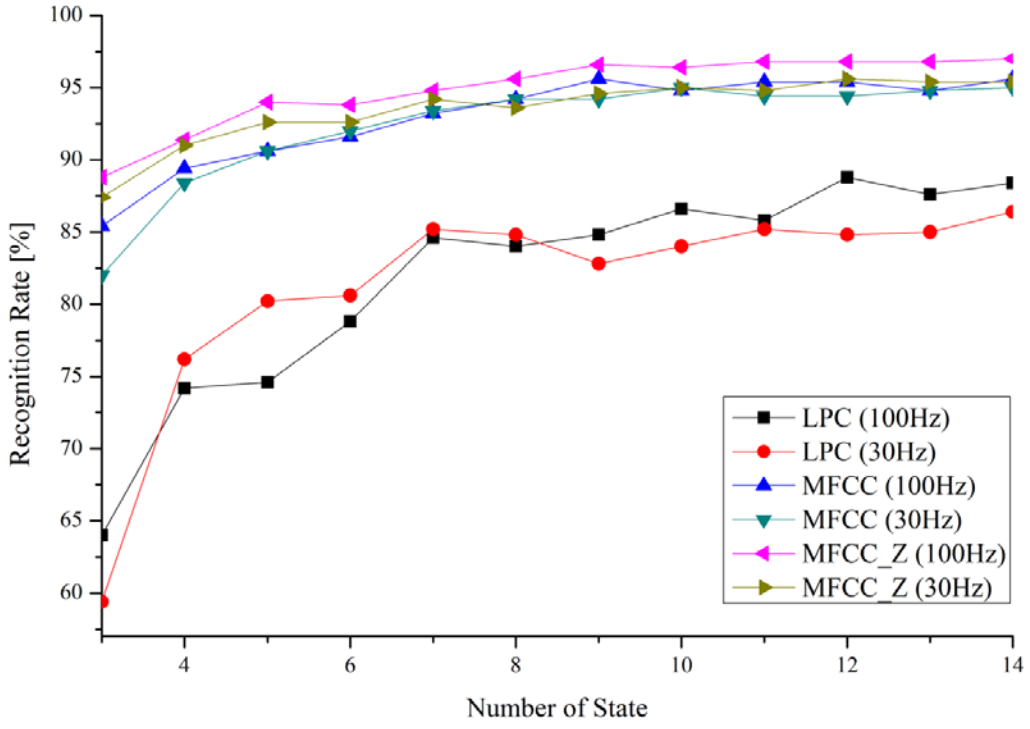


Fig.7.2: The effect of number of states on various feature types.

Tab. 7.14: Recognition rate [%] for isolated word speech recognition.

Number of states	LPC	MFCC	MFCC_Z	LPC	MFCC	MFCC_Z
	30Hz, 33.3333ms			*100Hz, 25ms*		
3	59.4	82	87.4	64	85.4	88.8
4	76.2	88.4	91	74.2	89.4	91.4
5	80.2	90.6	92.6	74.6	90.6	94
6	80.6	92	92.6	78.8	91.6	93.8
7	85.2	93.4	94.2	84.6	93.2	94.8
8	84.8	94.2	93.6	84	94.2	95.6
9	82.8	94.2	94.6	84.8	95.6	96.6
10	84	95	95	86.6	94.8	96.4
11	85.2	94.4	94.8	85.8	95.4	96.8
12	84.8	94.4	95.6	88.8	95.4	96.8
13	85	94.8	95.4	87.6	94.8	96.8
14	86.4	95	95.4	88.4	95.6	**97**

7.3 Experiments on LVCSR of Vietnamese

For experiments on LVCSR of Vietnamese, the speech corpus described in section 6.2 is used for training and testing purpose. Tab. 7.15 shows some statistics of this speech data. MFCC is the main feature type used for all experiments in which the speech signals are parameterized to generate a MFCC feature vector every 10ms using window size of 25ms. Each feature vector has 39 dimensions (12 MFCC coefficients, 1 energy coefficient and their first and second derivatives) and is applied CMN to further improve the recognizers. The HTK toolkit is the main tool used for estimating and evaluating the recognizers in all LVCSR tasks.

Tab. 7.15: Speech corpus for LVCSR tasks.

	Number of speakers		Number of utterances	Duration
	Male	*Female*		
Train	65	116	22,665	45 hours, 14 minutes
Test	4	6	535	1 hours, 25 minutes
Total	69	122	23,200	46 hours, 39 minutes

7.3.1 Examining the Effect of Tone in Vietnamese Syllables

To totally examine the effect of position and role of tone in the relationship with other components of a syllable, context-independent (monophone-based) continuous speech recognizers will be trained and tested using the first three strategies described in section 5.3. For each method in these strategies, the phonetic units are trained with 3 states HMM (not including start state and end state) using flat-start procedure. Each state of the phonetic

unit HMM consists of 8 Gaussian mixtures. The bi-gram LM is trained using the *VN only* part of the total text corpus (Tab. 6.9) in which the system's vocabulary contains all 5741 distinct syllables occurring in the transcription of the training utterances of the speech corpus. Turing-Good smoothing algorithm is used when training LM (Tab. 7.4).

Tab. 7.16 shows the syllable accuracy for all the methods of analyzing tone. From the recognition results, some major conclusions can be made by analyzing this information:

- Using tone versus without using tone: Because Vietnamese is a tonal language, the first problem when dealing with ASR of Vietnamese is to answer the question does tone has to be modeled to better describe syllables. Tab. 7.16 shows that for the same speech recognition strategy as well as phoneme set type, the analyzing methods which use tone give better result than the one without using tone. In the phoneme-based strategy, only two methods *C1wVTC2_I* and *C1wVTC2T_I* give worse result than the method without using tone *C1wVC2* (the reason will be explained in the next conclusions), the other methods all give better results. In the case of the vowel-based and rhyme-based strategies, it is obvious that methods with tone give better results. These results show that tone is an important component of a Vietnamese syllable and has to be modeled one or another way to obtain optimized results for LVCSR tasks.

- Independent tone versus dependent tone: So, if the tone has to be modeled, what is it role in the syllable? From the results, it is also easy to see that for the same analyzing method, dependent tone based methods always give better results than independent tone based methods. This conclusion can be explained by examining Fig.5.1 to Fig.5.3. Note that because the pitch information is not located at a specific position or area in the syllable but tends to spread along the syllable, especially on the rhyme part of the syllable, for methods with independent tone (*C1wVC2T_I*, *C1wVTC2_I*, *C1wVTC2T_I*, *C1MCT_I*, *C1MTC_I* and *C1RT_I*), the tone is supposed to locate after main vowel or at the end of syllable and captures only a small part of the pitch information incident to the speech signal, and so the tone cannot be fully modeled using this pitch information resulted in not good recognition results. For the other methods with dependent tone (*C1wVC2T_D*, *C1wVTC2_D*, *C1wVTC2T_D*, *C1MCT_D*, *C1MTC_D* and *C1RT_D*), by integrating itself into other components of the syllable (vowel, final component of syllable or both), the tone can capture pitch information on a larger area of the speech signal resulted in a better model

of tone can be constructed in the form of tonal phonetic unit, and a higher recognition result can be obtained. Another interesting aspect can also be seen is that, for dependent tone based methods, the method of rhyme-based strategy (*C1RT_D*) gives better result than other methods of vowel-based and phoneme-based strategies and the methods of vowel-based strategy give better results than the corresponding methods of phoneme-based strategy. This can be explained by having a look at Fig.5.1 to Fig.5.3. The pitch information covered by rhyme is usually larger than the one covered by vowel or phoneme. The same is also true in the case of vowel in comparison with phoneme.

The above information leads to the fact that tone cannot be modeled independently in the same manner with other phonetic unit. This problem can be solved by using one of the two methods: multi-streams HMMs and context-dependent (triphone-based) HMMs. For the first method, tone will be modeled independently with other phonetic units in a separate stream. The feature vectors that are used to model tone can be pitch or other parameter types that bear tonal information. Other phonetic units are modeled in a different stream to tone. This method also arises some difficult tasks that need to be solved such as tone corpus creation, selection of parameter type to best represent tone, fusion of tonal stream and phonetic unit's stream, streams synchronization, etc. For the second method, tone will be modeled independently in the same manner with other phonetic units, but by creating the relation between tone and other phonetic units in the form of context-dependent HMMs, tone is better modeled and the recognizer can better deal with tone. This method will be examined in the section 7.3.2.

- Tone position: The task of examining the position of tone in Vietnamese syllable is one of the most interesting and also challenging tasks. From Tab. 7.16, a very interesting property of tone in syllable is also presented. In the same speech recognition strategy as well as phoneme set type, the methods where tone is located at the end of syllable give better result than the methods where tone is located after main vowel both with dependent and independent tone hypotheses. For the phoneme-based strategy, methods *C1wVC2T_D* and *C1wVC2T_I* give better results than methods *C1wVTC2_D* and *C1wVTC2_I* respectively. It is also true for the vowel-based strategy in which methods *C1MCT_D* and *C1MCT_I* give better results than methods *C1MTC_D* and *C1MTC_I* respectively. The results above show that the important part of tone (part that carries the most pitch

information) is located at the end of syllable. This fact is true for Vietnamese. By changing the tone (length, height, etc.) of a syllable at the end of a pronouncing process, the meaning of a syllable can be changed totally. Together with the previous conclusions, this conclusion may help to explain the reason why the methods *C1wVTC2T_I* and *C1wVTC2_I* give worse result than the method *C1wVC2*. Although the method *C1wVTC2T_I* uses two separate tones (after main vowel and at the end of syllable) to describe a syllable, it supposes that these two tones present the same phonetic unit (Fig.5.1) which lead to inappropriate tone modeling. Note that, the method *C1wVTC2T_D* also uses two tones to describe a syllable, but these tones are modeled as two different tonal phonetic units resulted in a better representation of tone in syllable. In the case of the method *C1wVTC2_I,* tone occupies only a small part of pitch information located after main vowel (Fig.5.1), and because at this area in the syllable the pitch is not changed significantly (Fig.3.2), discriminative power of the six tonal models is small resulted in low SACC.

- Phone-based versus grapheme-based phoneme set: As mentioned in the previous chapters, two types of phoneme set can be used for LVCSR of Vietnamese tasks. Results from this experiment show that the analyzing methods based on these two phoneme sets are comparable. It is interesting to see that the grapheme-based phoneme set produces a little better results comparing to the phone-based phoneme set, especially for vowel-based and rhyme-based strategy. In general, each phoneme set has its own strong and weak points. For the grapheme-based phoneme set, the advantages are the simplicity in creating pronunciation dictionary where the phoneme is the grapheme itself, and the solution for the ambiguity of vowel phone /a/ where different sounds can be produced but presented by only one phone in the phone-based set. But the disadvantage is the ambiguity of consonant groups ‘ng’ – ‘ngh’, ‘g’ – ‘gh’, ‘k’ – ‘q’ – ‘c’; vowel group ‘i’ – ‘y’; and diphthong groups ‘yê’ – ‘iê’ – ‘ya’ – ‘ia’, ‘ươ’ – ‘ưa’, ‘uô’ – ‘ua’ where each group is presented by a single phone only (Tab. 5.2). On the other hand, the phone-based phoneme set is more suitable for hypothesis about Vietnamese phonology described in section 3.2. It overcomes the disadvantage of the grapheme-based phoneme set as well as makes the tasks of speech recognition of Vietnamese more understandable and applicable. The biggest problem of this method is that there are several hypotheses about the Vietnamese phonology (this thesis using the most popular one) with their own right and wrong aspects. The creation of

pronunciation dictionary which based solely on the grapheme-to-phoneme mapping table cannot be applied in some rare special syllables and the effect of dialects are also factors that affect the phone-based phoneme set. Solving these problems, which is not impossible, will definitely make the phone-based phoneme set becoming the most appropriate method to deal with speech recognition of Vietnamese.

- Strategies for LVCSR of Vietnamese: A final conclusion that can be made is the effect of various speech recognition strategies. The results from Tab. 7.16 show that the rhyme-based strategy gives better result than the vowel-based strategy and both strategies outperform the phoneme-based strategy. These results can be seen in two aspects. First, the rhyme component of syllable in the rhyme-based strategy and the vowel component of syllable in the vowel-based strategy are formed from one or more phonemes. As a result these phonetic units can be modeled in the form of monophone, diphone or triphone in comparison with other phonetic units in the phoneme-based strategy that modeled as monophone only. This makes the phonetic units of rhyme-based and vowel-based strategies tend to carry more information than other phonetic units of the phoneme-based strategy both in the sense of longer speech signal information and the relation between phonemes. Second, for methods which using dependent tone, by attaching into the rhyme or vowel, the tone (in the form of tonal phonetic unit) can capture a larger part of pitch information in the syllable (Fig.5.1 to Fig.5.3), and so tone can be better modeled for the recognizers.

Tab. 7.16: SACC [%] for context-independent LVCSR.

Dictionary type		**Phoneme set type**	
		Phone-based	*Grapheme-based*
Phoneme-based strategy	*C1wVC2*	45.56	45.93
	C1wVC2T_D	**59.53**	**59.24**
	C1wVC2T_I	47.51	48.08
	C1wVTC2_D	52.96	54.08
	C1wVTC2_I	43.66	44.68
	C1wVTC2T_D	58.28	58.42
	C1wVTC2T_I	41.89	42.12
Vowel-based strategy	*C1MC*	49.82	50.80
	C1MCT_D	**63.34**	**63.01**
	C1MCT_I	50.96	51.71
	C1MTC_D	57.90	58.18
	C1MTC_I	50.81	52.27
Rhyme-based strategy	*C1R*	58.51	58.60
	C1RT_D	**65.05**	**65.26**
	C1RT_I	59.93	60.18

7.3.2 Context-dependent LVCSR of Vietnamese

As mentioned in the previous sections, the phoneme-based strategy, which based solely on monophone, can be further improved by applying context-dependent HMM. In this experiment, the performance of this type of recognizer will be examined using various methods in the phoneme-based strategy. For each method, the phonetic units are first trained with 3 states HMM using flat-start procedure. Then a set of context-dependent syllable internal triphone acoustic models are trained in which similar acoustic states of these triphones are tied using tree-based clustering method. Each state of the phonetic unit HMM consists of 8 Gaussian mixtures. For this experiment, the same LM described in section 7.3.1 will be used for training and testing purpose.

The syllable accuracies of all methods presented in the phoneme-based strategy are shown in Tab. 7.17. In the table, it is easy to see that context-dependent HMM does improve the SACC of recognizers trained using phoneme-based strategy. The interesting aspect is that all methods in this strategy outperform the best method *C1RT_D* of the context-independent HMM-based recognizers described in the previous experiment.

Another conclusion, which reaffirms the previous conclusion about tone, is that tone has to be used to better model the Vietnamese syllables. All the methods with tone modeling in the phoneme-based strategy give better results than the method without using tone (*C1wVC2*), and the best result obtained when using method *C1wVC2T_I*. This also verifies the conclusions in the previous section that the most important part of tone information is located at the end of syllable and tone can be modeled independently when the relationship between tone and other phonetic units is constructed using context-dependent HMM. Again, the phone-based and grapheme-based phoneme sets are comparable to each other.

Tab. 7.17: SACC [%] for context-dependent LVCSR.

Dictionary type	**Phoneme set type**	
	Phone-based	*Grapheme-based*
C1wVC2	68.52	68.60
C1wVC2T_D	72.07	71.94
C1wVC2T_I	**73.90**	**73.77**
C1wVTC2_D	71.48	71.81
C1wVTC2_I	72.87	72.66
C1wVTC2T_D	71.50	71.56
C1wVTC2T_I	71.36	71.26

7.3.3 The effect of LM on LVCSR of Vietnamese

In these experiments, the performance of recognizers in LVCSR task will be examined using the best method *C1wVC2T_I* in the context-dependent scheme. The phone-based phonetic units are first trained with 3 states HMM using flat-start procedure. Then a set of context-dependent syllable internal triphone acoustic models are trained in which similar acoustic states of these triphones are tied using tree-based clustering method. Each state of the phonetic unit HMM consists of 8 Gaussian mixtures.

7.3.3.1 The effect of text corpus category

For this experiment, the bi-gram LMs are trained using the *VN only* set of the text corpora described in section 6.1.4 in which the system's vocabulary contains all 5741 distinct syllables occurring in the transcription of the training utterances of the speech corpus. Good-Turing smoothing is used when training LMs. The perplexities of the LMs are shown in Tab. 7.1 to Tab. 7.4.

Because the testing speech data contains text only in the literature domain, some conclusions can be seen from the syllable accuracies shown in Tab. 7.18. First, LM estimated from text corpus of type *News* gives the worst SACC than other types. It is a reasonable result given the type of text this LM covering with the type of text in the testing data. On the other hand, the LM constructed from the general purpose text corpus provides really good result in comparison with LM constructed from text corpus of type *Literature*. This shows that the text corpus collected using the method described in section 6.1.2 is good enough for general purpose because it covers text in many fields such as news, literature, law, etc. Also, as expected, recognizer using LM of *Literature* category obtains really good SACC. It is interested to see that the combined text corpus of the three categories above gives the best SACC.

Tab. 7.18: SACC [%] for various text corpus categories.

LM type	SACC [%]
Literature	73.17
News	69.90
General purpose (GP)	72.31
Literature + news + GP	**73.90**

7.3.3.2 The effect of dictionary size and smoothing method

For this experiment, all the bi-gram LMs of three different smoothing methods including Good-Turning, Kneser-Ney and Witten-Bell as well as various vocabulary sizes will be examined (Tab. 7.5 to Tab. 7.8).

Tab. 7.20 shows the syllable accuracies of all the recognizers. For the smoothing methods, recognizers based on the Kneser-Ney smoothing using interpolated model give the best SACCs, and both of the backoff and interpolated model of the Kneser-Ney smoothing give better SACCs in comparison with Good-Turning and Witten-Bell smoothing.

It is also easy to see that, for all of the smoothing methods, LMs with vocabulary containing all syllables occurring in the text corpus give the best SACCs and the LMs with vocabulary size of 5741 syllables provide the worst results. Despite this fact, the SACCs resulted from recognizers that use these LMs are not different much, especially when the vocabulary size is larger than or equal to 6000 syllables. This can be explained by analyzing Tab. 7.19. This table presents the total percent of text in the training text corpus covered by the syllables in the vocabulary. It shows that even with the vocabulary of 5741 syllables, it covers more than 99% of text in the text corpus. This is a very interesting aspect where LM is based on syllable and it verifies the fact that the most common Vietnamese syllables are from 7000 to 8000.

Tab. 7.19: Percent of text covered by syllables in vocabulary.

Vocabulary size	Covered text [%]	Covered syllable
5741	99.718	71,534,926
6000	99.926	71,683,737
7000	99.968	71,713,650
All	100	71,736,930

Tab. 7.20: SACC [%] for various smoothing method and vocabulary size.

Smoothing method	Vocabulary size			
	5741	*6000*	*7000*	*All*
Good-Turning	73.90	73.92	73.92	**73.95**
Kneser-Ney	73.89	73.96	**73.97**	73.95
Kneser-Ney interpolation	74.03	74.08	74.10	**74.11**
Witten-Bell	73.68	73.75	73.77	**73.79**

7.3.3.3 The effect of multi-syllable-based LM

For this experiment, multi-syllable-based LMs described in section 7.1.2 are used. To estimate and evaluate the recognizers using this type of LM, all the training and testing transcription of the speech corpus will be segmented using procedure 7.1. An interested aspect of multi-syllable-based recognizer is that internal triphone set is formed not only from inside a syllable but also between syllables of a bi-syllabic token. This makes a broader dependent context will be modeled.

Tab. 7.21 shows the syllable accuracies of all the recognizers trained using different values of bi-syllabic token *N* to segment text for LM and acoustic model estimation. It is easy to see that, the larger the number of bi-syllabic tokens used for segmenting text the better the recognition result will be. When *N* is increased to 40000, the best SACC is obtained. At this point, keep increasing *N* do not improve the SACC. The results also show that the recognizers using multi-syllable-based LM outperform the case where syllable-based LM is utilized. An improvement of nearly 5% of the SACC is obtained using this type of LM (the best SACC of 78.70% resulted from multi-syllable-based LM in comparison with 74.11% of syllable-based LM). There are two main reasons for this result. First, by concatenating syllables into multi-syllabic tokens, more history is put into LM where bi-gram LM contains the context of not only two neighbor syllables but also three or four neighbor syllables. Second, the context-dependent acoustic models are modeled both within a syllable and across syllables in the form of bi-syllabic tokens which reduce the confusion when recognizing short syllables.

Tab. 7.21: SACC [%] for multi-syllable-based LMs.

N	SACC [%]
50	73.87
80	74.02
100	74.21
500	74.70
1,000	74.92
5,000	74.93
10,000	75.80
15,000	77.57
20,000	77.84
30,000	78.36
40,000	**78.70**
50,000	78.43

7.3.4 Gender-dependent LVCSR of Vietnamese

In all the previous experiments on LVCSR of Vietnamese, speaker-independent strategy was used to train all the recognizers. For this strategy, the variability of speaker's attributes is undesirable and unavoidable in which speaker's gender is one of the most influential sources of this variability. So, by developing a gender recognizer, the performance of speech recognition systems can be further improved both in the stability and accuracy. In this work, a simple gender recognizer will be constructed using Gaussian Mixture Model (GMM) on MFCC feature. The using of GMM as a representation of speaker identity is motivated by: First, each component of GMM may model some underlying set of acoustic classes which reflect some general speaker-dependent vocal tract configurations that are useful for characterizing speaker identity. Second, a linear combination of Gaussian basis functions is capable of representing a large class of sample distributions (the ability to form smooth approximations to arbitrarily-shaped densities).

To train and test the gender recognizer, a speech data of 240 speakers (120 male speakers and 120 female speakers) is collected. This data is divided into two parts: the training part (100 male speakers and 100 female speakers) and the testing part (20 male speakers and 20 female speakers). All the speech data will be converted to MFCCs feature vectors and modeled by GMMs. In this experiment, the number of mixture of GMMs is changed from 1 to 84 to find the best number of mixture. The results show that the gender recognition rate is 100% when the number of mixture larger or equal to 7.

For gender-dependent LVCSR task, all the phone-based recognizers are trained using method *C1wVC2T_I* in the context-dependent scheme as described in the previous experiments. The recognizers for male and female speaker are estimated separately using the above speech corpus. In this corpus, 8360 utterances (15 hours, 27 minutes) of male speakers and 14305 utterances (29 hours, 47 minutes) of female speakers are used as training data, and the other 297 utterance (41 minutes) of male speakers and 238 utterances (44 minutes) of female speakers are used as testing data. Both of the syllable-based and multi-syllable-based bi-gram LMs using Kneser-Ney smoothing with interpolated model will be examined in this experiment. For multi-syllable-based LM, N = 40000 is used for segmenting the training and testing data.

Tab. 7.22 shows the syllable accuracy of gender-dependent recognizers using different types of LM. It is easy to see that the gender-dependent recognizers do improve the recognition results for all type of LM in which multi-syllable-based LM gives the best result. The best SACC of 79.66% is an improvement in comparison with 74.11% of gender-independent recognizers with syllable-based LM.

Tab. 7.22: SACC [%] for gender-dependent recognizers.

		SACC [%]		
		Male	*Female*	*All*
Syllable-based	*Gender dependent*	74.00	77.23	75.63
	Gender independent	x	x	74.11
Multi-syllable based	*Gender dependent*	79.14	80.17	**79.66**
	Gender independent	x	x	78.70

Chapter 8

Audio-Visual Speech Recognition of Vietnamese

8.1 Introduction

Visual information has shown to be a useful source to improve the performance of ASR systems in noise conditions [140-142]. Studies on the field of lipreading and human perception motivated many works that take advantage of the visual information in speech recognition tasks, ranging from isolated word [143] to connected word or continuous speech recognition tasks [144-146]. The performance gains of audio only ASR in noisy condition is noticeable when integrated with visual channel.

The main issue in all of the audio-visual ASR systems is the designing of a visual front end procedure to extract features containing important speech information from a sequence of speaker's face in a given video. Each visual front end has to deal with the following problems: 1) face and facial features localization, 2) region of interest (ROI) extraction, 3) visual features type selection, and 4) visual features improvement.

For the first two problems, various methods have been proposed in the literature. In [145, 147], color-based segmentation and Fisher discriminant are employed to perform face detection and facial feature localization, then the mouth region is extracted using feature points obtained from these methods. This approach tends to find only points at specific locations on face such as pupils, nostrils, corner of mouth, etc., and so, could not be used to extract the exact boundaries of mouth. To overcome this, model-based methods have been applied with many degree of success. Among them, shape-based model [148, 149] and appearance-based model [146, 150, 151] were widely used in lipreading field.

To deal with the third problem, various visual feature types have been proposed which can be grouped into three categories: shape-based, appearance-based and the combination

of both. For shape-based features, the lip contours are usually extracted and their properties such as width, height, perimeter, moments, etc., are utilized as visual features. The appearance-based features have been proved to outperform the shape-based features [152] and the extraction of these features is obtained on the basis of image transform techniques such as DCT, DWT, DFT, PCA, etc. In the third category, model based approach such as active appearance model (AAM) [146, 151] is used to obtain visual features in the form of model's parameters.

In the last problem, the improvement of visual features can be obtained on the image level or on the feature level. For image level improvement, one can change the size of ROI to cover different parts of the face region [142], normalize the extracted ROI by removing differences in size and rotation between frames or make the number of image frames for each utterance equal [153]. For feature level improvement, feature mean normalization (FMN) can be applied to account for variation in lighting condition. The classification power of visual feature and also the dimensional reduction can be further improved by using LDA and maximum likelihood linear transform (MLLT) both in frame level and across frames level to incorporate dynamic information into the visual features.

In this work, a robust and efficient method call constrained local models (CLM) [154-156] is used to extract face and facial feature boundaries. With the tracking power of CLM, shape-based feature (mouth and face boundaries) as well as appearance-based feature (DCT, PCA, DWT, etc.) and model-based feature (AAM) can be fully extracted. Among these features, DCT, PCA, and AAM are chosen as the base features for experiments on audio-visual speech recognition because they provide good visual features and also there are fast algorithms to compute these features. The visual feature is further improved in term of discriminating power and dimensional reduction by using LDA based data projection. The effect of interpolation, number of feature selection, type of visual front end and the method of LDA training will also be examined.

8.2 Feature extraction

8.2.1 Face and facial features localization

In computer vision, the task of face and facial feature localization is always a challenging task. To deal with this problem, many approaches have been presented in literatures with

varying degrees of success. In this work, a method called CLM, which is an extension of discriminative method for face and facial features alignment and tracking, has been proposed as a promising approach that can accomplish the given task. The basic structure of CLM is shown in Fig.8.1. In CLM, a point distribution model (PDM) and a patch model are first built and then combined in the fitting procedure to locate face and facial feature in an image or a video frame. The process of face and facial features alignment consists of two steps: first, a quick search using a well-known technique called Viola-Jones face detector is applied to locate the face location in the image in form of rectangle surrounding the face. Then, a more detail scan on the returned rectangle face location using CLM is applied resulting in the final facial feature boundaries. Note that, when tracking video frame for face, only the first frame needs the quick searching step, other frames just need the second step to do face alignment.

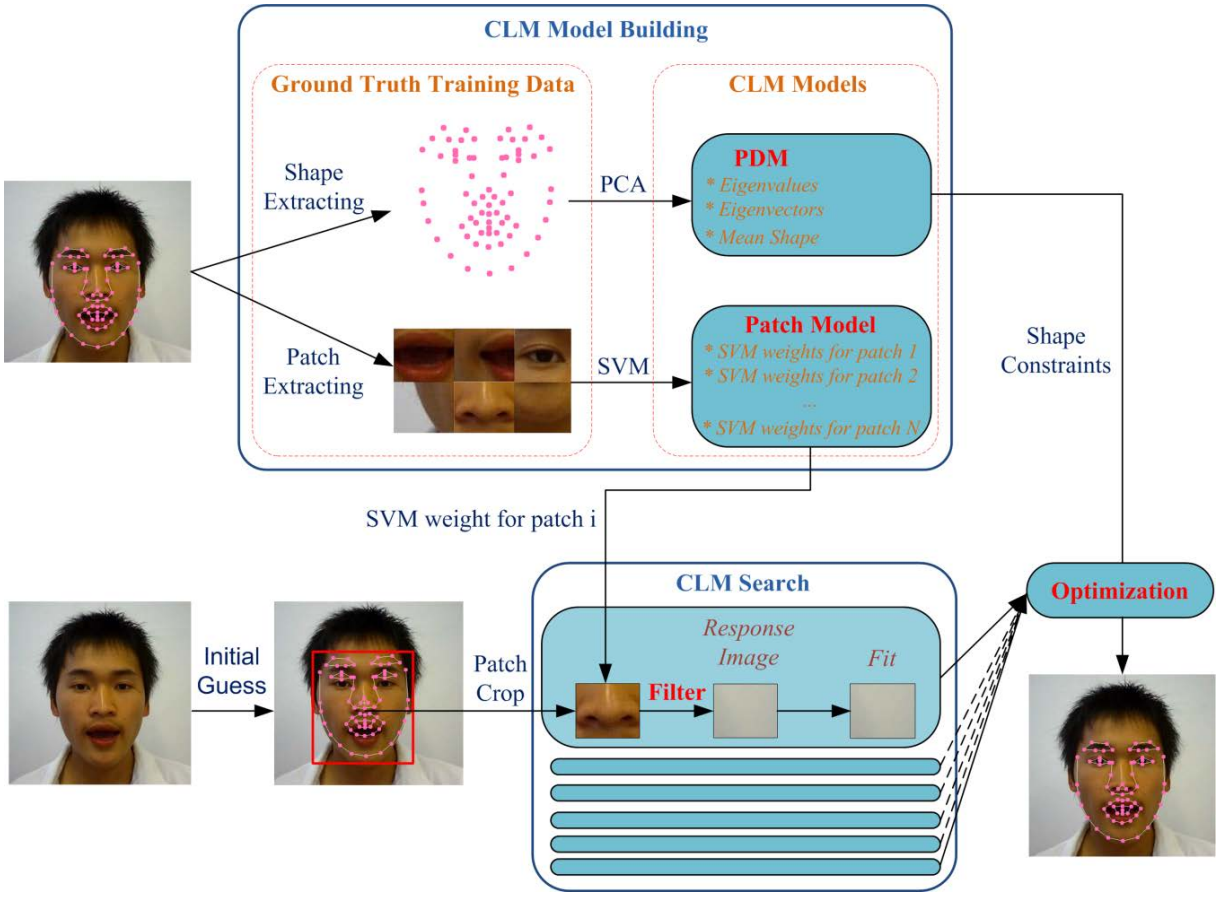


Fig.8.1: The Architecture of CLM.

8.2.2 Region of interest extraction

The process of extracting ROI of mouth can be seen on Fig.8.2. Having obtained the optimized shape of the current frame from the previous step, the task of ROI extracting becomes very simple. First, a ROI of size 64 x 64 pixels around the center point of mouth is selected. Then, to remove the differences in scale and rotation of ROI across frames of the same video and across videos, the face in this video frame is resized according to the same reference shape before the ROI is extracted. The size of reference shape and optimized shape are computed using pixel as unit (e.g. width of shape is 60 pixels, height of shape is 80 pixels).

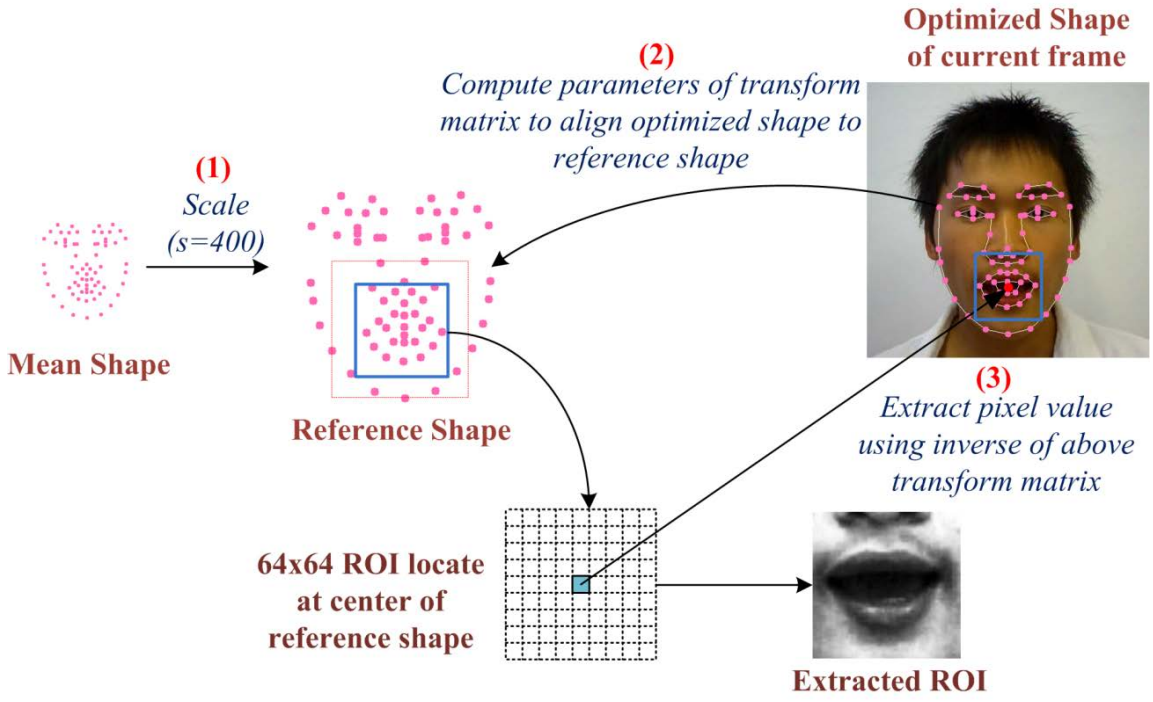


Fig.8.2: ROI extraction.

Note that, when we say the ROI of size *NxN* pixel, we do not know exactly how much the ROI will cover the mouth region (which parts of the mouth and face are inside ROI) supposing that center of ROI locate at center of mouth. To determine how much the *NxN* ROI will cover the mouth region, reference shape will be used as the base reference. By changing the size of reference shape, the region covered by *NxN* ROI will be different (Fig.8.2 shows that the red rectangle and blue rectangle cover different region of mouth of reference shape when changing the reference shape size). The reference shape can be obtained in many ways. One possible way is to use the mean shape of PDM obtained when

training CLM models. This mean shape will have unit length. In the experiments, the mean shape is scaled 400 times to obtain reference shape in which the 64x64 ROI will cover the main part of the mouth as can be seen in the extracted ROI in the figure below.

8.2.3 LDA data projection

Given a set of class $C = [C_1, ..., C_K]$ and a data matrix $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{d \times N}$ consisting of N vectors $\{\mathbf{x}_i\}_{i=1}^{N}$ in $\Re^d$ which are labeled as $C(i) \in C$. LDA find a projection matrix **P** to map a data point $\mathbf{x}_i$ in the d-dimensional space to a data point $\mathbf{y}_i$ in the lower l-dimensional space as follows:

$$\mathbf{x}_i \in \Re^d \rightarrow \mathbf{y}_i = \mathbf{P}^T \mathbf{x}_i \in \Re^l (l < d).$$

To train matrix **P**, two matrices called within-class scatter matrix $\mathbf{S}_W$ and between-class scatter matrix $\mathbf{S}_B$ are defined:

$$\mathbf{S}_W = \frac{1}{N} \sum_{i=1}^{K} \sum_{x \in C_i} (\mathbf{x} - m_i)(\mathbf{x} - m_i)^T . \qquad (8.1)$$

$$\mathbf{S}_B = \frac{1}{N} \sum_{i=1}^{K} N_i (m_i - m)(m_i - m)^T . \qquad (8.2)$$

where m_i is the mean of the class C_i, m is the global mean, N_i is the sample size of class C_i, and $\sum_{i=1}^{K} Ni = N$.

Matrix **P** is then estimated by computing the generalized eigenvalues $\mathbf{\Lambda}$ and right eigenvectors **V** of the matrix pair $(\mathbf{S}_B, \mathbf{S}_W)$ that satisfies $\mathbf{S}_B \mathbf{V} = \mathbf{S}_W \mathbf{V} \mathbf{\Lambda}$. The first l eigenvectors of **V** corresponding to l largest eigenvalues will form the projection matrix **P**.

In this work, inner frame and across frames LDA will be trained. For inner frame LDA, visual feature vector extracted from each video frame will be considered as a sample to train LDA matrix. Using this type of LDA features, we can select not only the first few highest energy visual coefficients but also some other types of coefficients which capture more useful information from ROI of mouth. In the case of across frames LDA, the feature vector center at current frame will be concatenated with its neighbor frames to capture temporal visual information before fed to the training process. Note that the temporal feature vector can be formed from the output of coefficients extracting stage or from the output of inner frame LDA extracting stage.

The selection of a set of phonetic class when training LDA will affect the discriminating power of the LDA matrix. In this work, the phonetic classes are selected based on the analysis of Vietnamese syllable. Tab. 8.1 shows the methods which will be used to train LDA matrix and their corresponding number of phonetic classes in the training data. Because it is difficult to label phoneme boundaries using video frames, the boundaries of C1, w, V and C2 in the audio part of the audio-visual database are manually labeled and then the visual frame will be assigned to the corresponding audio class (phoneme, rhyme, di-phone, etc.) using the labeled data. Fig. 8.3 and Fig. 8.4 show how a visual frame is assigned to its corresponding audio frame. You can see that for temporal visual frames located at the start and at the end of an audio frame, it contains not only visual information of the same class but also of its neighbor class.

Tab. 8.1: Analyzing of Vietnamese syllabel into basic units.

Method	Syllable Analysis Description	No. of class
C1wVC2	Four phonemes: C1, w, V, C2	41
C1wVC2T	Four phonemes, tone is attached to the last phoneme of syllable: C1, w, V, C2, T	155
C1wVTC2	Four phonemes, tone is attached to V: C1, w, VT, C2	111
C1R	Phoneme and rhyme: C1, wVC2	151
C1RT	Phoneme and rhyme, tone is attached to rhyme: C1, wVCT	449
DD	Tri-phone and bi-phone: C1wV, VC2	461
DDT	Tri-phone and bi-phone, tone is attached to bi-phone: C1wV, VC2T	730
DTDT	Tri-phone and bi-phone, tone is attached to both: C1wVT, VC2T	1367

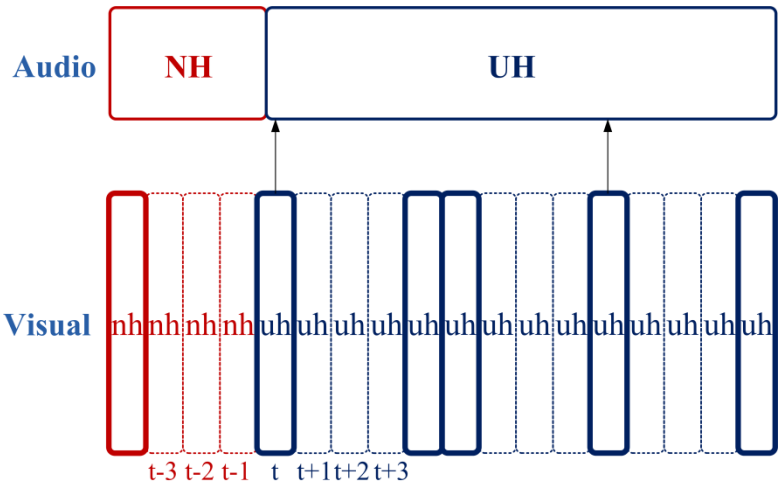


Fig. 8.3 : Across frame features selection.

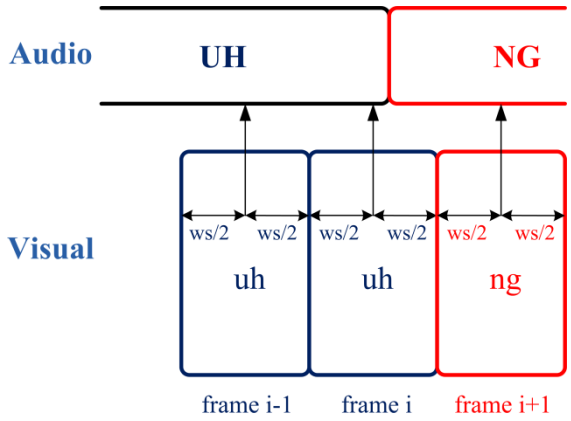


Fig. 8.4: Assigning of feature vectors to the audio classes.

8.2.4 Visual front end for feature extraction

Given the ROI of mouth and the optimized face boundaries of video frames, the visual features can now be extracted using DCT, PCA or AAM (Fig.8.5). For DCT features, 2D DCT is applied to the grayscale ROI resulted in a 2D coefficient matrix. Then the first D1 coefficients are extracted in zigzag pattern and stored in a visual feature vector. To extract PCA features, a vector of pixel value is formed by concatenating all columns of the grayscale ROI. PCA matrix is then applied to this concatenated vector resulted in a vector of PCA coefficients, and the first *D1* coefficients corresponding to the first largest eigenvalues are extracted as visual feature vector. The AAM features will be extracted using two types of model: AAM model of mouth and AAM model of face. These models are applied to the video frame using the optimized face of the current frame to obtain the vector of model parameters. Again, the first *D1* parameters corresponding to the first largest eigenvalues are extracted and stored as a visual feature vector. The visual frame will be examined at sample rates of 30Hz and 100Hz, and so for sample rate 100Hz, an interpolating step is needed because the video was sampled at 30Hz. This feature vector is further processed by mean of feature mean normalization (subtract this vector to the mean vector computed over the frame of the current video). In this work, two types of visual front end can be applied to improve the classification power of the final feature vector: 1-Stage LDA visual front end and hierarchical LDA (HLDA) visual front end. For 1-Stage LDA, the output feature vector from the FMN steps is concatenated with its neighbor feature vectors center at current frame and then this concatenated vector is multiplied with the LDA matrix to obtain the final feature vector. For hierarchical LDA visual front end,

the feature vector is first projected to a new vector space using inner frame LDA, then across frame LDA will project the feature vector output from the first stage to obtain the final visual feature vector.

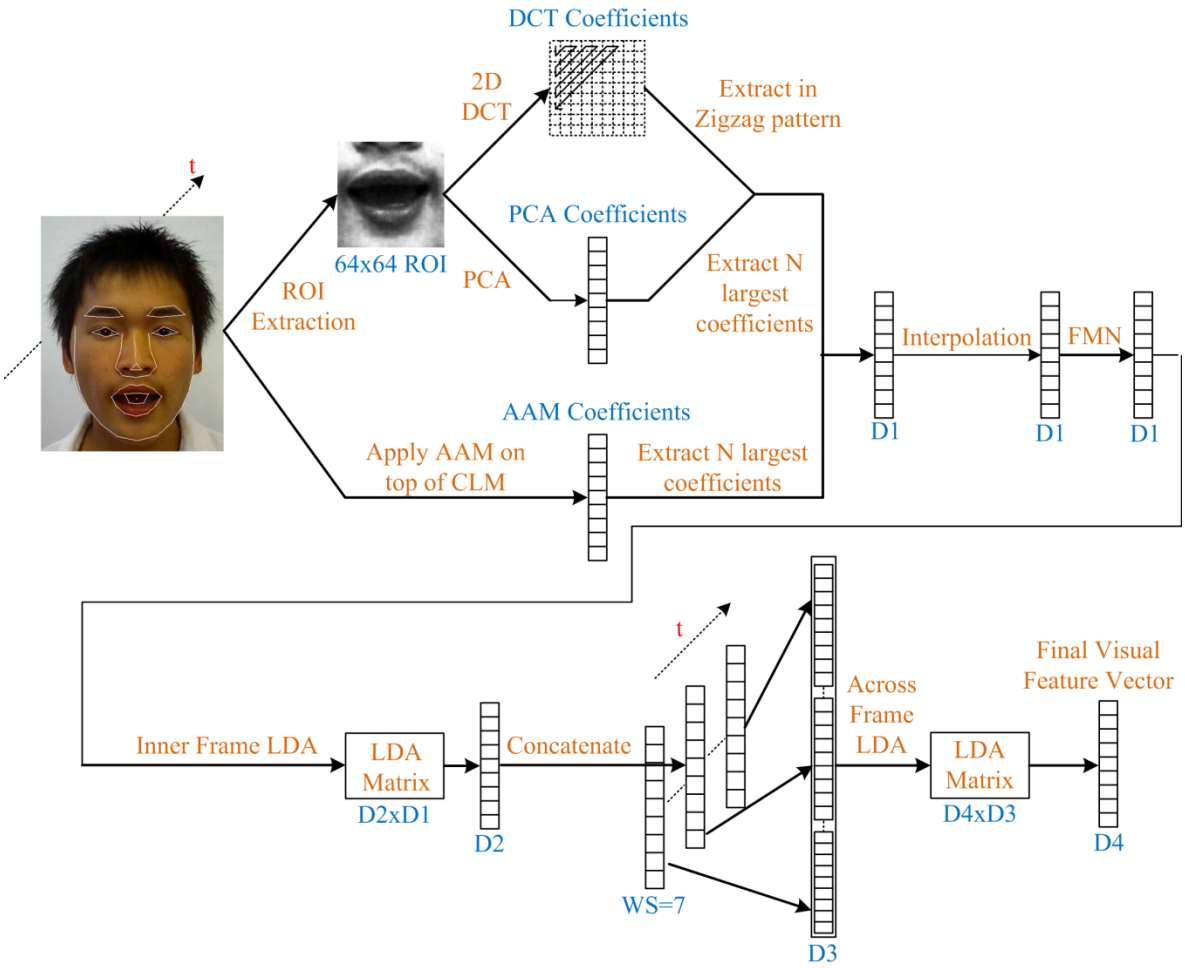


Fig.8.5: Visual front end feature extraction.

8.3 Isolated word visual only speech recognition

In these experiments, all the isolated words are modeled by HMMs using the HTK toolkit. The isolated word part of the audio-visual database is divided into two groups: the first group containing 40 speakers is used for training all HMMs, and the second group containing 10 speakers is used for evaluating of isolated word speech recognition task. To trained LDA matrix, the continuous speech part of the audio-visual database (2000 specific sentences and 600 general sentences uttered by 40 speakers) is manually labelled in phoneme level using the Praat toolkit.

For experiments in this section, DCT visual feature is used as the main feature to examine two types of visual front end: HLDA and 1-Stage LDA. In the first experiment, the effect of the number of DCT coefficients *D* on inner frame LDA is examined. Other parameters such as sampling rate *F* and the size of feature vector output from inner frame LDA that obtaining highest accuracy *dmax* are also studied to obtain the best parameter set. The LDA matrices (trained using method C1wVC2) are used to extract visual feature for the isolated word data. Then 14-states HMMs of all isolated words are trained and the testing results are shown in Tab. 8.2. It is easy to see that larger number of DCT coefficient results in better accuracy. This result means that less significant DCT coefficients still hold useful information of ROI of mouth.

Tab. 8.2: Recognition results (VI) for various visual parameters using inner frame LDA.

D	dmax	F (Hz)	VI [%]
50	10	30	54
100	6	30	55.8
200	10	30	**56.6**
50	14	100	50
100	13	100	51.6
200	10	100	53.6

In the second experiment, the effect of different sets of basic phonetic class on inner frame LDA is examined. The best number of DCT coefficient resulted from previous step (*D* = 200) will be used in this experiment. Tab. 8.3 shows that for Vietnamese, training LDA matrix using phoneme as basic class with tone attached to the last component of the syllable (C1wVC2T) give the best result both in accuracy (57%) and dimensional reduction (8). From both experiments, we can also see that interpolation the visual stream from sample rate 30Hz to 100Hz has degraded the recognition results.

In the third experiment, the performance of two visual front ends for feature extraction is compared. For this experiment, the best parameters result from previous experiments (*D* = 200, *dmax* = 8, *F* = 30Hz) are fed to HLDA visual front end and for 1-Stage LDA the parameters will be *D* = 50, *F* = 30Hz. Both visual front ends will be examined with different window sizes (*WS*) and LDA matrix is trained using C1wVC2T method. Tab. 8.4 shows the best recognition results for each visual front end. For both method, the best accuracy is obtained at *WS* = 7. It can also be seen from Fig.8.6, the HLDA visual front end

outperform the 1-Stage LDA visual front end both in the highest accuracy and average accuracy and both visual front ends do improve the DCT only visual feature.

Tab. 8.3: Recognition results for various methods of Vietnamese syllable analysis using inner frame LDA (VI).

Method	30Hz		100Hz	
	dmax	*VI (%)*	*dmax*	*VI (%)*
C1wVC2	10	56.6	12	53.6
C1wVC2T	**8**	**57**	8	53.2
C1wVTC2	12	56.4	12	53
C1R	13	56.2	16	50.8
C1RT	9	55.6	16	50.8
DD	15	55.4	8	51
DDT	14	56.4	14	52.8
DTDT	9	56	13	51.4

Tab. 8.4: Recognition results using HLDA (VH) and 1-Stage LDA (VS) with different WS.

WS	HLDA		1-Stage LDA	
	dmax	*VH (%)*	*dmax*	*VS (%)*
7	16	**62**	15	58.4
8	10	61.6	16	57.4
9	10	60.8	20	57.4
10	10	61.2	20	58
11	10	61.4	20	57.4

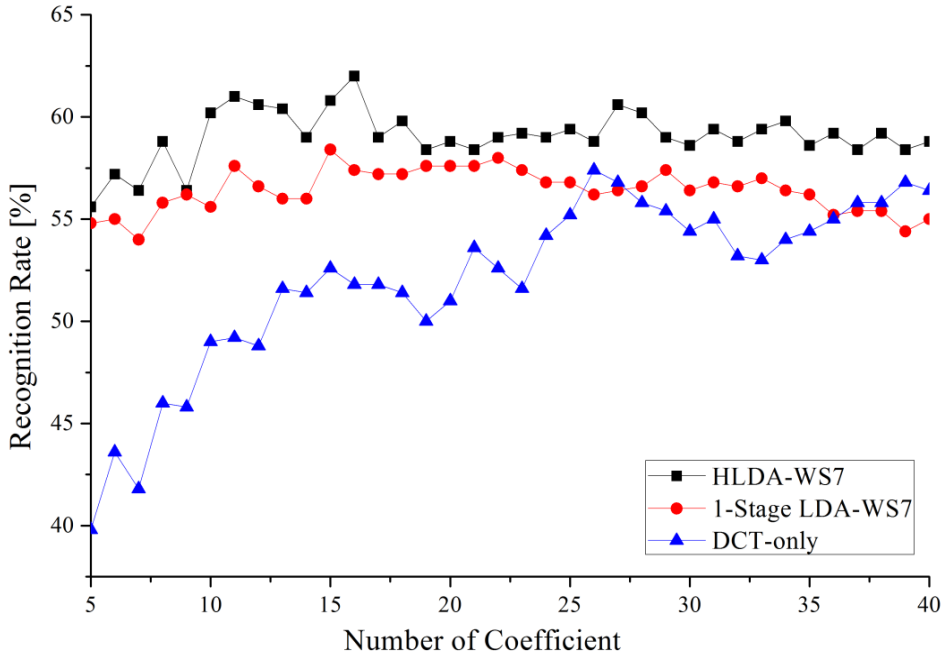


Fig.8.6: HLDA and 1-Stage LDA for DCT coefficient with WS = 7.

In this experiment, three types of visual feature (DCT, PCA and AAM) are compared using the best visual front end and parameters obtained from the previous experiments. The HLDA visual front end is used with sampling rate F = 30Hz and LDA training method C1wVC2T as the base system to extract these types of visual feature. Tab. 8.5 shows the effect of HLDA at different WSs. The best accuracy is from DCT and PCA (62%), but DCT obtains better dimensional reduction (16). Fig.8.7 also shows that, DCT and PCA feature outperform the AAM feature when using HLDA as basic visual front end.

Tab. 8.5: Recognition Results Using HLDA with Different Types of Visual Feature.

WS	DCT		PCA		AAMLip		AAMFace	
	dmax	*VH (%)*	*dmax*	*VH (%)*	*dmax*	*VH (%)*	*dmax*	*VH (%)*
7	**16**	**62**	40	**62**	19	56.4	31	57.8
8	10	61.6	40	61.2	21	56.6	33	58
9	10	60.8	34	61.2	33	57	35	57.6
10	10	61.2	36	61	33	57.2	34	58.2
11	10	61.4	25	60.8	23	57.8	32	59.2

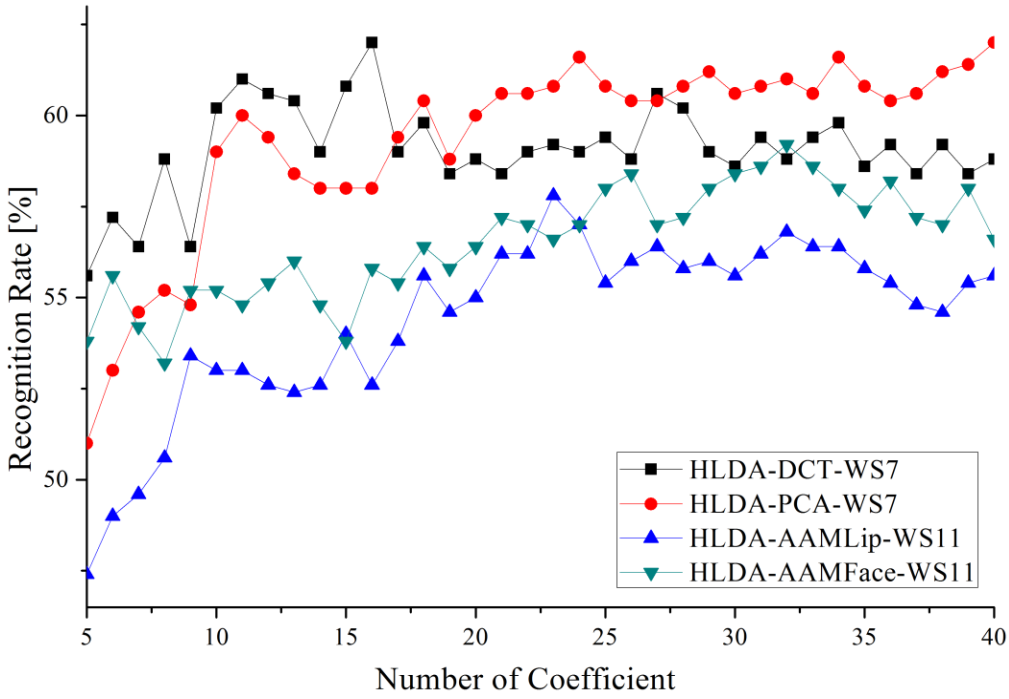


Fig.8.7: HLDA for DCT, PCA and AAM using the best WS for each feature type.

8.4 Isolated word audio-visual speech recognition.

8.4.1 Audio-visual integration

The audio and visual stream can be integrated in different ways corresponding to one of the three levels of integration: Early integration (EI), middle integration (MI) and late

integration (LI). These integration methods require the computation of a likelihood. Considering a task of classifying an utterance O given by,

$$\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T\} \qquad (8.3)$$

where T is the number of observation and $\mathbf{o}_T$ denotes the feature vector for observation t. The likelihood $p(\mathbf{O} \mid \boldsymbol{\lambda}_i)$ for HMM parametric class model $\boldsymbol{\lambda}_i$ given the utterance $\mathbf{O}$ is found canonically by expanding all possible state paths,

$$p(\mathbf{O} \mid \boldsymbol{\lambda}_i) = \sum_{\text{all } \mathbf{q}} p(\mathbf{O}, \mathbf{q} \mid \boldsymbol{\lambda}_i) \qquad (8.4)$$

This full likelihood is estimated in practice using the Viterbi approximation which only requires a single path,

$$\log p(\mathbf{O} \mid \boldsymbol{\lambda}_i) \approx \frac{1}{T} \log p(\mathbf{O}, \mathbf{q}^* \mid \boldsymbol{\lambda}_i) \qquad (8.5)$$

where $\mathbf{q}^* = \{q^*(1), ..., q^*(T)\}$ is the optimal state path. An a posteriori probability estimate for modality $m \in \{a, v, av\}$, referring to the acoustic, visual and audio-visual modalities respectively, can be found through Bayes rule, assuming equal priors, from the estimated likelihoods such that,

$$\hat{P}r(\omega_i \mid \mathbf{O}^{\{m\}}) = \frac{p(\mathbf{O}^{\{m\}} \mid \boldsymbol{\lambda}_i^{\{m\}})}{\sum_{n=1}^{N} p(\mathbf{O}^{\{m\}} \mid \boldsymbol{\lambda}_n^{\{m\}})} \qquad (8.6)$$

Both the EI and MI strategies employ an audio-visual HMM during evaluation so equation 7.6 can be used directly to obtain an audio-visual confidence score $\zeta(\omega_i \mid \mathbf{O}^{\{av\}}) = \hat{P}r(\omega_i \mid \mathbf{O}^{\{av\}})$. The EI strategy obtains its audio-visual HMM through training directly from synchronized audio-visual features. The MI strategy employs independently trained HMMs from the acoustic and visual modalities from which a composite audio-visual HMM is created known as a multi-stream HMM with the output state probabilities given by:

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} \mathrm{N}(o_{st}; \mu_{jsm}, \sum_{jsm}) \right]^{\gamma_s} \qquad (8.7)$$

where M_s denote the number of mixture components in streams, c_{jsm} is the weight of the *m*-th component and *N* is a multivariate Gaussian with mean vector μ_{jsm} and covariance matrix $\sum_{jsm}$ and γ_s is a stream weight.

The LI strategy uses two independently trained HMMs from the acoustic and visual modalities for evaluation, and so it cannot employ equation 8.6 directly to obtain an audio-visual confidence score. The confidence score for the recognition process can be express as a function of each modality's a posteriori probability estimate,

$$\zeta(\omega_i \mid \mathbf{O}^{\{av\}}) = F(\hat{P}r(\omega_i \mid \mathbf{O}^{\{a\}}), \hat{P}r(\omega_i \mid \mathbf{O}^{\{v\}})) \qquad (8.8)$$

In LI strategy, a given acoustic and visual features of an unknown class can be classified into an utterance class ω^* using,

$$\omega^* = \arg\max_i \{\gamma \log P(\mathbf{O}^{\{a\}} \mid \boldsymbol{\lambda}_i^{\{a\}}) + (1-\gamma) \log P(\mathbf{O}^{\{v\}} \mid \boldsymbol{\lambda}_i^{\{v\}})\} \qquad (8.9)$$

where $\boldsymbol{\lambda}_i^{\{a\}}$ and $\boldsymbol{\lambda}_i^{\{v\}}$ are the acoustic and visual HMMs of *i*-th class, $\log P(\mathbf{O}^{\{a\}} \mid \boldsymbol{\lambda}_i^{\{a\}})$ and $\log P(\mathbf{O}^{\{v\}} \mid \boldsymbol{\lambda}_i^{\{v\}})$ are their output log-likelihoods respectively. The integration weight γ determines how much the final decision relatively depends on each modality. It has a value between 0 and 1, and varies according to the amounts of noise contained in speech.

From equation 8.9, one can choose a constant weight value over various noise conditions or exhaustedly search for the best weight γ in a certain noise condition. One can also select the weight as a function of the signal to noise ratio (SNR) by assuming that the SNR of the acoustic signal is known (this assumption is not always feasible). Another way to determine the weight is to use an adaptation data.

The best method in practice is to automatically adjust the weight without the need of a priori knowledge of the current noise condition or an additional adaptation data. This way, when the acoustic speech is clean, the weight should be large because recognition of clean acoustic speech usually outperforms that of the visual speech. On the other hand, when the acoustic speech contains a lot of noise, the weight should be sufficiently small. One of the most popular methods among such schemes is the reliability ratio-based method in which the weight is calculated from the relative reliability measures of the two modalities. The reliability of each modality can be measured from the outputs of the corresponding HMMs.

When the acoustic speech does not contain any noise, there are large differences between the acoustic HMMs' outputs. The differences become small when the acoustic speech contains noise, which reflects increased ambiguity in recognition due to the noise. The confidence can be defined in various ways:

The variance of log-likelihood outputs

$$\sigma = \frac{1}{N-1}\sum_{n=1}^{N}(R_n - \overline{R})^2 \qquad (8.10)$$

where R_n is the n-th output of the classifier corresponding to N-best hypotheses in each modality and $\overline{R}$ is the mean of N output.

The N-best ratio average

$$\sigma = \frac{1}{N-1}\sum_{n=2}^{N}(R_1 - R_n) \qquad (8.11)$$

where R_n is equal to the n-th best hypotheses and all R are sorted in descending order, such that this is the difference between the best hypothesis and the rest.

The N-best dispersion

$$\sigma = \frac{2}{N(N-1)}\sum_{n=1}^{N}\sum_{n'=n+1}^{N}(R_n - R_{n'}) \qquad (8.12)$$

where $N \geq 2$ and R_n is equal to the n-th best hypotheses.

Obtaining the confident scores, the weighting factor γ is calculated using ratio-based method,

$$\gamma = \frac{w * \sigma_a}{\sigma_a + \sigma_v} \qquad (8.13)$$

where σ_a and σ_v are the confident scores for the output of the acoustic and visual stream, respectively and w is adjusting coefficient.

8.4.2 Audio-visual fusion experiments

For experiments on audio-visual fusion, all adaption and testing data will be synthesized in noise condition. To add noise to audio signal, the signal to noise ratio (SNR) is first estimated using the audio-visual database. For better capture dynamic changes of speech signal, segmental signal to noise ratio (SSRN) is used,

$$SSNR = 10.\log\left(\frac{1}{F}\sum_{j=0}^{M-1}\frac{\sum_{i=0}^{N-1}s_j^2[i]}{\sum_{i=0}^{N-1}n_j^2[i]}\right) \quad (8.14)$$

where F is the number of frames from speech signal and M is the length of one frame in sample. The resulted SSNR from the testing audio signal is about 29 dB.

Because the audio-visual database is created in room condition, SNR is unknown. So, we have to artificially add noise to audio signal using the algorithm described in [157]. In the next experiments, the noises contain in the NOISEX database [158] is used as the additive noise signals. First, all the training, adaption and testing data are processed and parameterized. The audio and visual features extracted from 40 speakers are used as training data for audio only, visual only and audio-visual isolated word speech recognition experiments. Features extracted from other 5 speakers are used as testing data for the recognition task, and audio-visual features extracted from the last 5 speakers are used as adapting data.

For audio-visual fusion experiments, the best parameters for the acoustic and visual feature obtained from previous experiments are used. Both acoustic and visual features are extracted at 30Hz and the resulted features are used to train word-based HMMs with 14 stages and 1 mixture. For audio stream, MFCC is used as acoustic feature with size of 39 dimensions (13 static coefficients with their 13 delta and 13 acceleration coefficients). For visual stream, DCT feature is used. From the previous experiments, HLDA matrix trained using C1wVC2T method is used to extract the final DCT visual feature with size of 16.

8.4.2.1 Middle integration experiments

For the experiments on MI of audio and visual stream, three types of noise are used as additive noises including white noise, babble noise and volvo noise. Fig.8.8 shows that in noise condition, the result for visual only speech recognition is 62.2% and does not change when noise change. But for audio only isolated word speech recognition, the recognition results reduce when the SSNR change from 29dB to -4dB. It can also be seen that different types of noise affect the audio stream in different degree, and in high noise condition the performance of audio stream degrade rapidly.

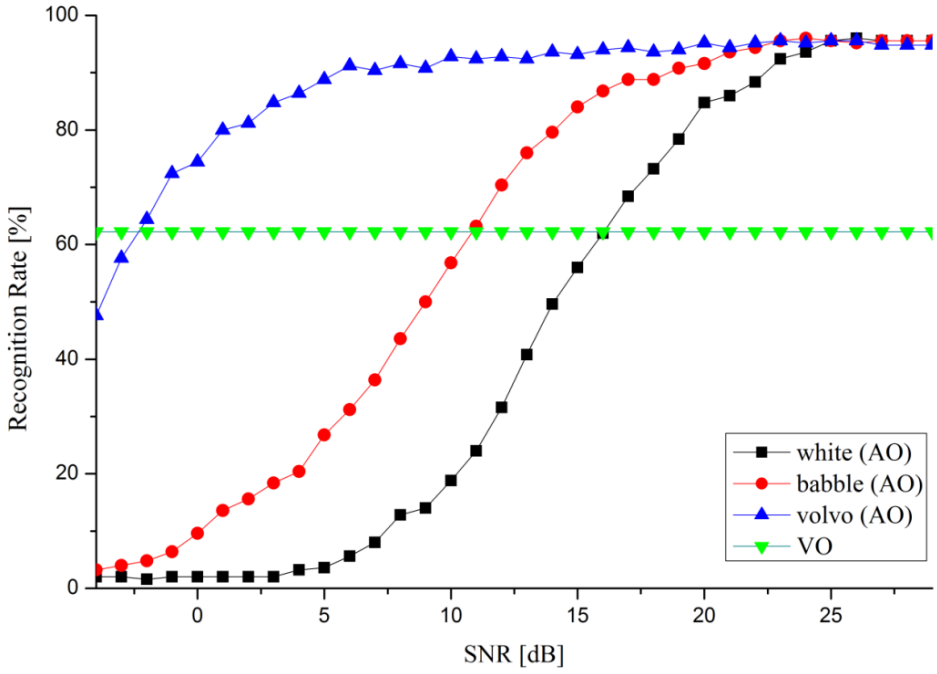


Fig.8.8: Recognition results for audio only (AO) and visual only (VO) using additive noises.

Fig.8.9 to Fig.8.11 show the results of audio-visual fusion using MI strategy. In these experiments, two-stream HMMs are used to train each isolated word in which the audio and visual stream weights are the same and equal to 1. It is easy to see that audio-visual fusion do improve the recognition results in noise conditions.

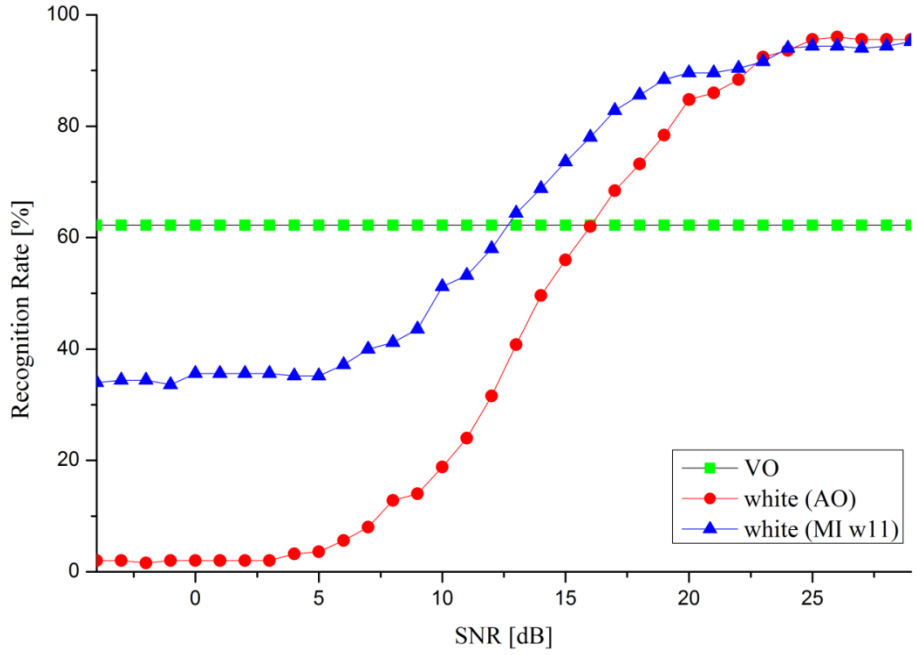


Fig.8.9: Recognition results for audio only (AO), visual only (VO) and audio-visual (MI) in white noise condition.

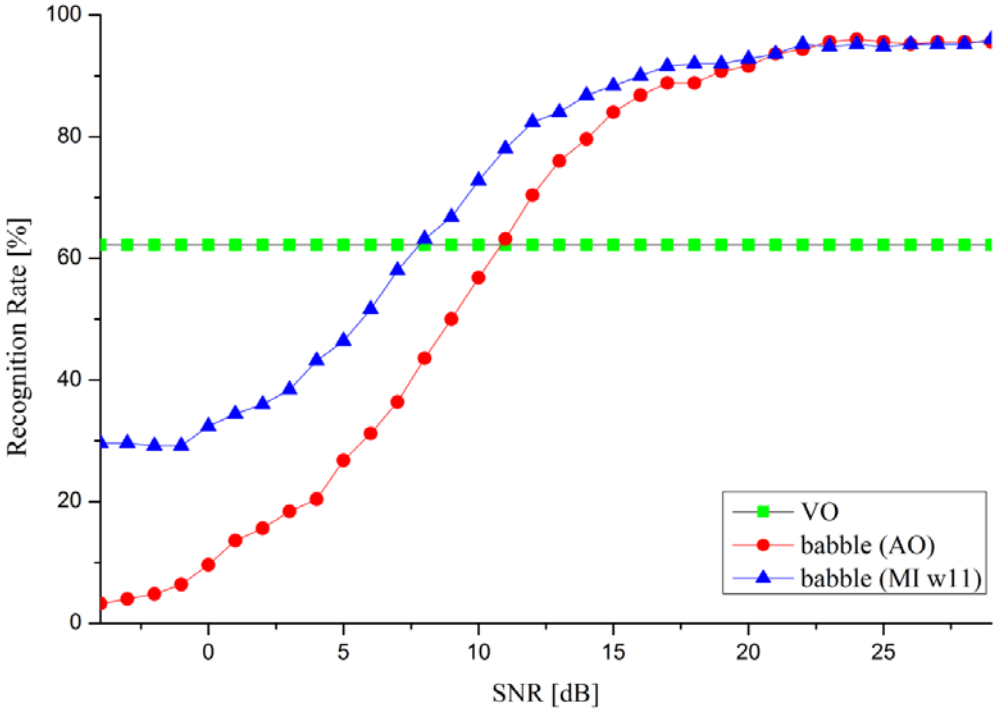


Fig.8.10: Recognition results for audio only (AO), visual only (VO) and audio-visual (MI) in babble noise condition.

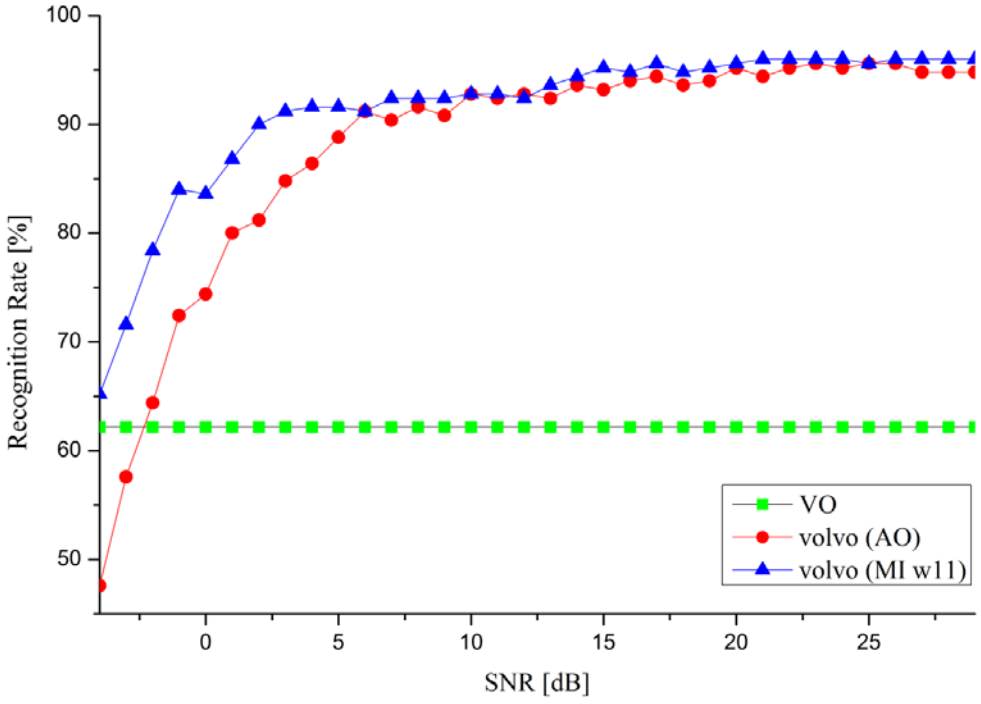


Fig.8.11: Recognition results for audio only (AO), visual only (VO) and audio-visual (MI) in volvo noise condition.

In the above experiments, both audio and visual stream weights are equal to 1 which are not optimal for different SSRNs. When SSNR is small (<10dB), visual stream tends to outperform audio stream and vice versa. To overcome this problem, an experiment to determine the best weights for each stream in different SSNR values is performed by changing the audio and visual stream weight from 0 to 3 with step 0.1. This experiment is done using adaptation data and the best stream weights at different SSNRs obtained from this step is used on the testing data. Fig.8.12 and Tab. 8.6 show that, for white noise, MI using adaptation data outperform the case where both stream weights equal to 1, especially in high noise condition (SSNR < 10dB).

The results from the previous experiments show that audio-visual speech recognition (AVSR) outperform audio only speech recognition for the task of isolated word recognition of Vietnamese especially when SSNR values are in the range from -4 to 20 dB. It can also be seen that white noise has the biggest effect on the decline of recognition result of audio only speech recognition. By using adapting data, MI strategy obtains optimal recognition results, especially in very high noise condition.

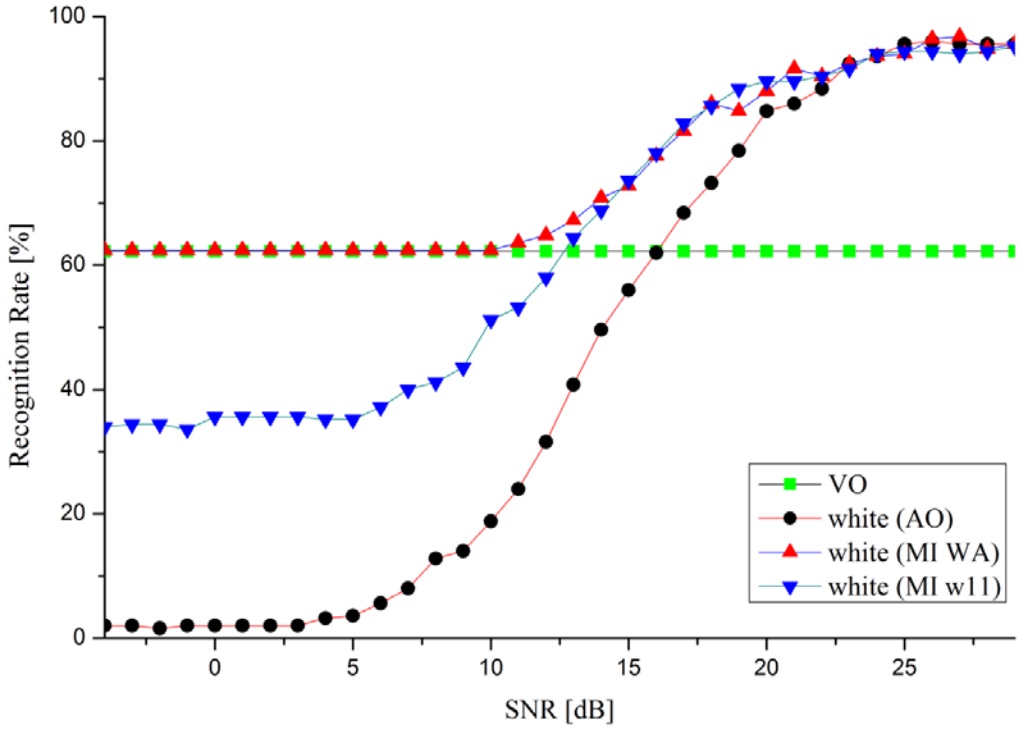


Fig.8.12: MI using adaptation data (WA) compare to equal weight (W11) in white noise condition.

Tab. 8.6: recognition results for MI using both stream weights = 1 and using the best stream weights for each SSNR.

SSNR (dB)	Audio weight	Visual weight	Recognition rate [%]	SSNR (dB)	Audio weight	Visual weight	Recognition rate [%]
-4	0	2.6	62.4	13	0.6	1.4	67.2
-3	0	2.6	62.4	14	0.4	0.8	70.8
-2	0	2.6	62.4	15	1.0	1.6	72.8
-1	0	2.6	62.4	16	1.4	2.0	77.6
0	0	2.6	62.4	17	1.0	1.6	81.6
1	0	2.6	62.4	18	1.6	2.2	86
2	0	2.6	62.4	19	1.0	2.2	84.8
3	0	2.6	62.4	20	1.4	2.6	88
4	0	2.6	62.4	21	2.2	1.6	91.6
5	0	2.6	62.4	22	1.4	1.6	90.4
6	0	2.6	62.4	23	0.8	1.0	92.4
7	0	2.6	62.4	24	2.6	3.0	93.6
8	0	2.6	62.4	25	1.4	1.6	94
9	0	2.6	62.4	26	0.6	0.2	96.4
10	0	2.6	62.4	27	0.8	0.2	**96.8**
11	0.4	1.4	63.6	28	0.8	1.0	94.8
12	0.6	1.8	64.8	29	0.8	1.0	95.6

8.4.2.2 Late integration experiments

In the first experiment, LI using exhausted search strategy is examined. For this strategy, the best audio weight γ for each SSNR value is first exhaustedly search using adapting data in which the audio weight γ is changed from 0 to 1 with step 0.05. The best audio weight γ obtained from this step for each SSNR value will be applied on the testing part of the database.

Fig.8.13 to Fig.8.15 show the recognition results for three different noises using this LI strategy. These figures also show the optimal recognition results (LI ES) where the exhausted search is applied directly on the testing data. This case can be considered as the optimal result of LI using exhausted search strategy. It is easy to see that the audio visual fusion using this LI strategy outperform the audio only and visual only in most of the noise condition. The results also show that using adapting data for audio weight searching can obtain the results as good as the optimal case.

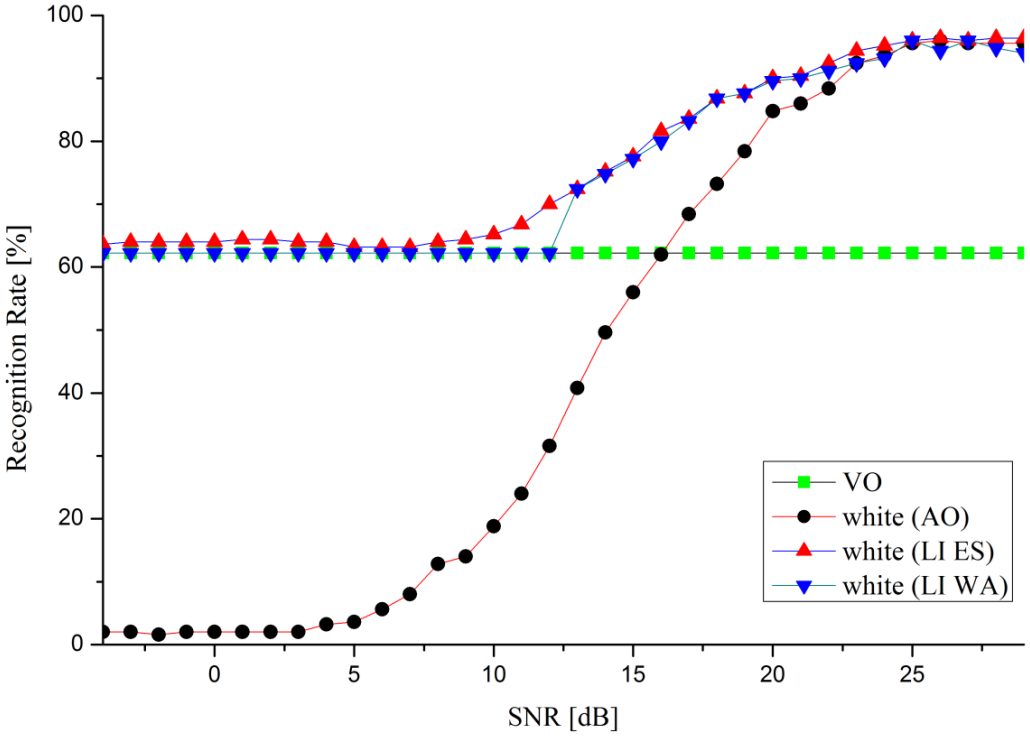


Fig.8.13: LI using exhausted search strategy (LI WA) in white noise condition.

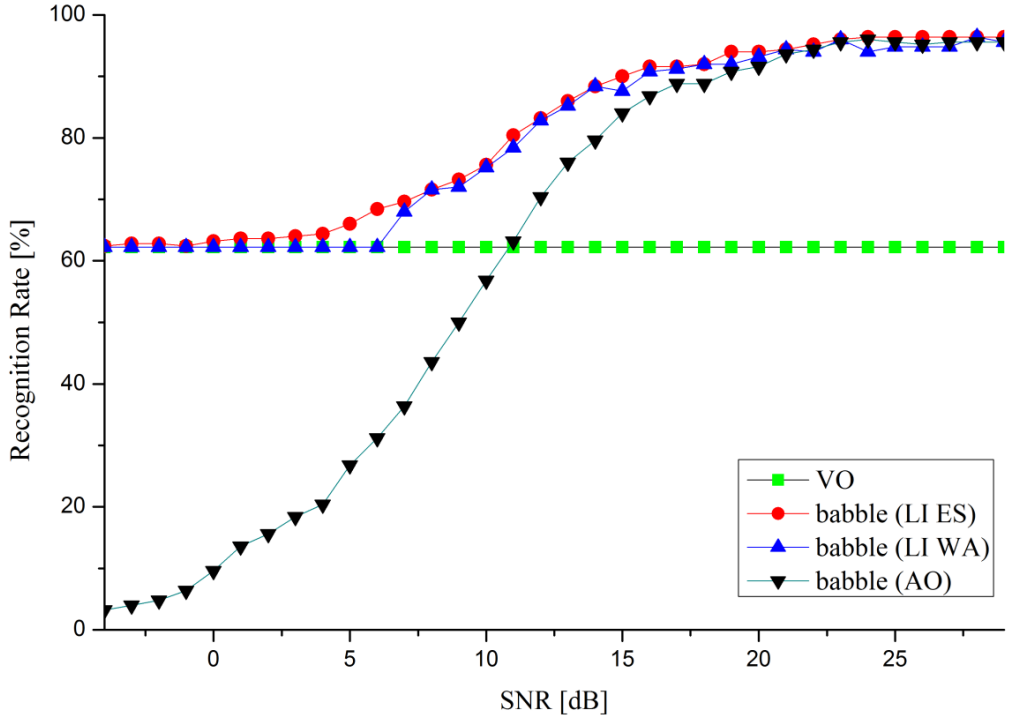


Fig.8.14: LI using exhausted search strategy (LI WA) in babble noise condition.

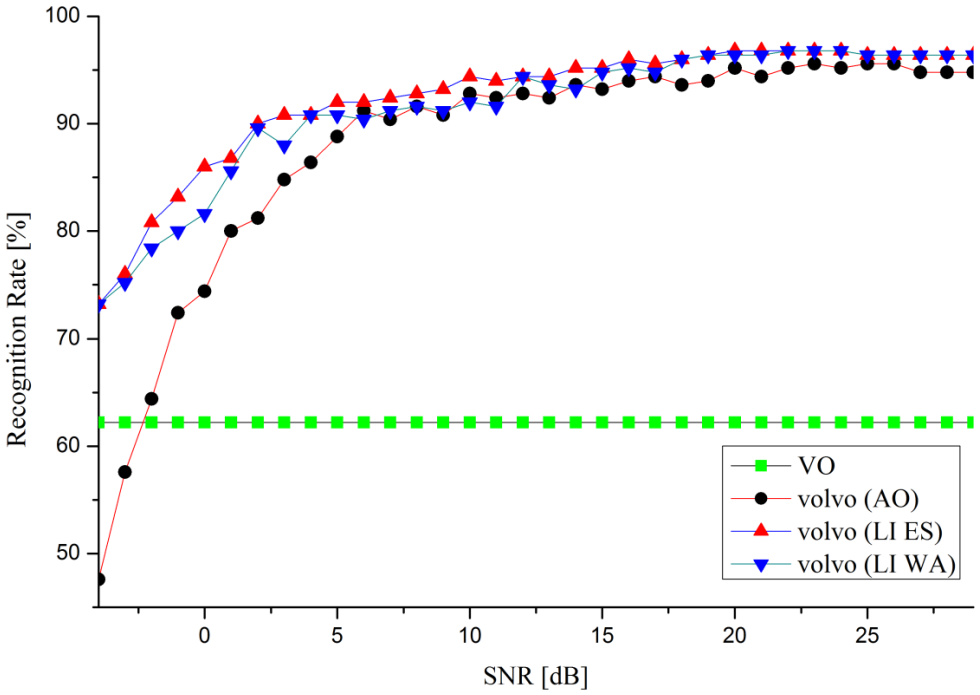


Fig.8.15: LI using exhausted search strategy (LI WA) volvo noise condition.

In the previous experiments, the audio weight is manually selected using adapting data which is difficult in practice due to lack of an available data or noise condition is unknown. In the next experiments, the audio weight is automatically computed using equation 8.13. The confidence scores in this equation are computed using three methods: variance (Var), N-best dispersion (Disp) and N-best ratio average (Aver). All of these methods compute the confidence scores using N-best hypotheses. Because the database contains 50 isolated words, each HMM will produce 50 hypotheses ranking from the highest score (log-likelihood) to the lowest score which will be used to compute the confidence for each auditory and visual stream.

To determine the number of N-best hypotheses used to compute the confident score for each methods in different noise conditions, the confidence scores are computed for each stream using N-best hypotheses ranging from 2 to 50 with SSNR value in the range from -4 to 29 and the adjusting weight w in the range from 0.05 to 1 (equation 8.13). Tab. 8.7 shows the N-best hypothesis for each method in three different noise conditions and their corresponding highest recognition rate. It shows that, using the first 9 to 12 best outputs (log-likelihood) of HMM is good enough to compute the confidence score for auditory and visual stream.

Tab. 8.7: N-best hypotheses for each confidence score type.

Method	N-best		
	White noise	*Babble noise*	*Volve noise*
Variance	9 (96%)	11 (96.4%)	11 (96.4%)
N-best dispersion	11 (95.6)	12 (96.4%)	11 (96.4%)
N-best average	11 (95.6)	9 (96%)	9 (96.4%)

To see the effect of adjusting weight w in equation 8.13, different values of w at the given N-best values obtained from the above experiment are examined.

Fig.B.1 to Fig.B.9 show that, for white noise and babble noise, in high noise condition (SSNR < 12), applying w with small value ($w \leq 0.5$) gives better results. Contrary to that, in clean or low noise condition (SSNR> 11), applying w with large value ($w > 0.5$) gives better results. This means that, in high noise condition, the auditory stream is declined, and so applying small value of w will reduce the importance of auditory weight and allow visual stream to be more efficient. On the other hand, in low noise or clean conditions, the auditory stream will be more efficient, so applying high value of w will increase the recognition rate of the system. Fig.8.16 and Fig.8.17 show a way to optimize the system in white noise and babble noise using confidence score as LI strategy. By applying small weight (w = 0.05) in high noise conditions and large weight (w = 1) in low noise or clean conditions, the recognizer using optimal weight w to compute confident scores for LI (CS Opt) outperforms the recognizer using only weight w = 1 to compute confident scores (CS w1).

For volvo noise, it seems that large values of weight w (w > 0.85) is good enough to obtain high recognition results even in the case of high noise condition. The reason is that the effect of volvo noise on the auditory stream is not too much. Fig.8.18 shows the effect of LI using w = 1 in volvo noise condition.

Fig.8.19 to Fig.8.21 compare the performance of isolated word recognition using MI strategy, LI strategy with audio only (AO) and visual only (VO) in noise conditions. It is easy to see that the fusion of audio and visual using MI or LI outperform the audio only for speech recognition task in which the LI perform better in some aspects: recognition rate, runtime stream weight modification, the independence in the combination of two streams output (log-likelihood), etc., and so it makes LI more applicable than MI.

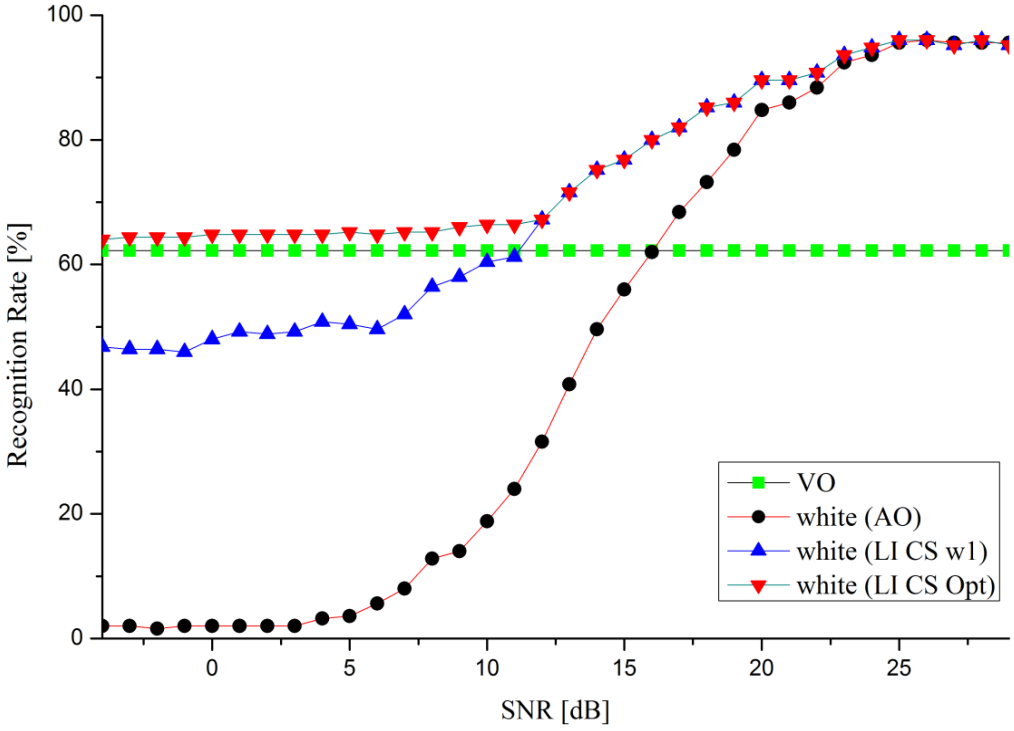


Fig.8.16: LI using confidence score strategy in white noise.

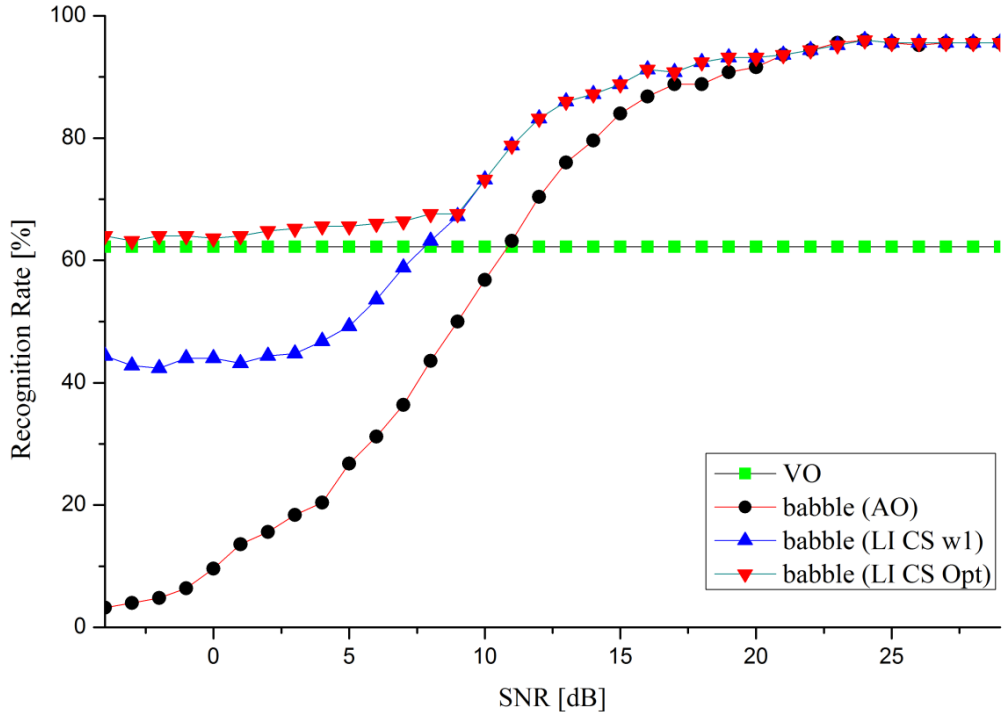


Fig.8.17: LI using confidence score strategy in babble noise.

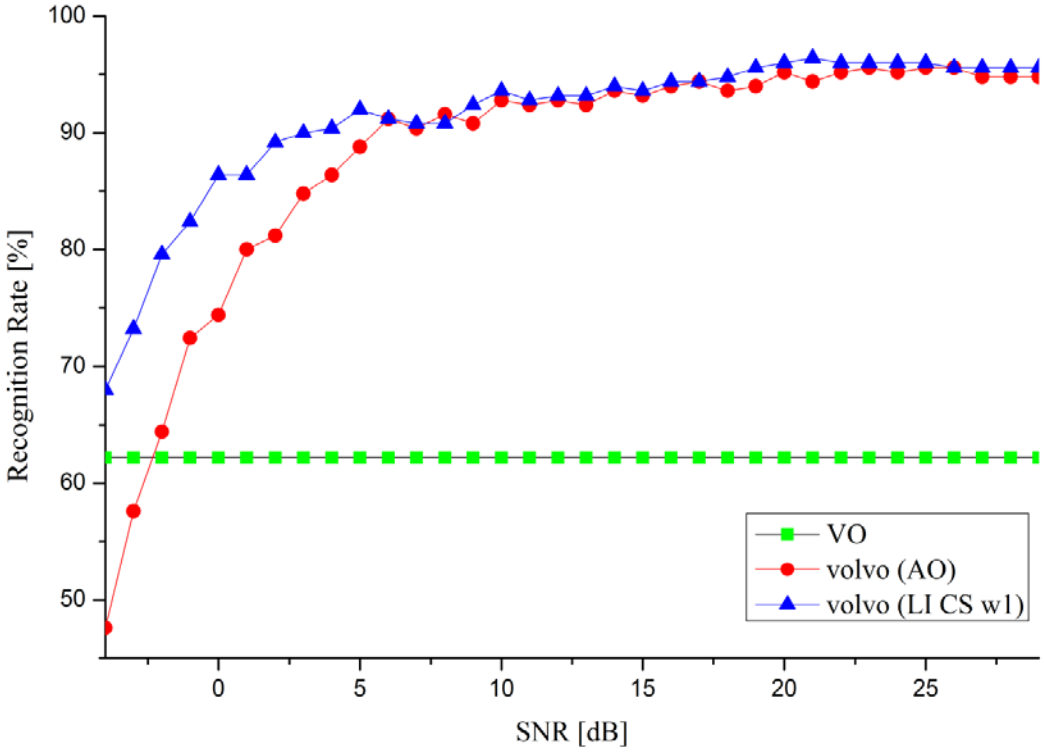


Fig.8.18: LI using confidence score strategy in volve noise condition.

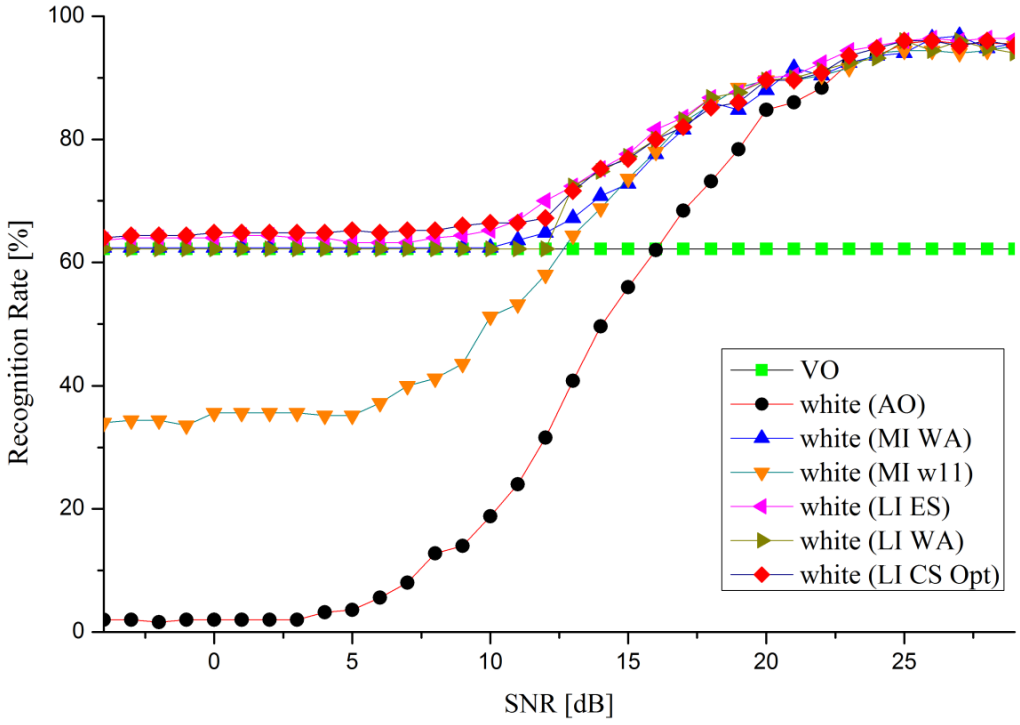


Fig.8.19: Comparison of fusion strategies in white noise.

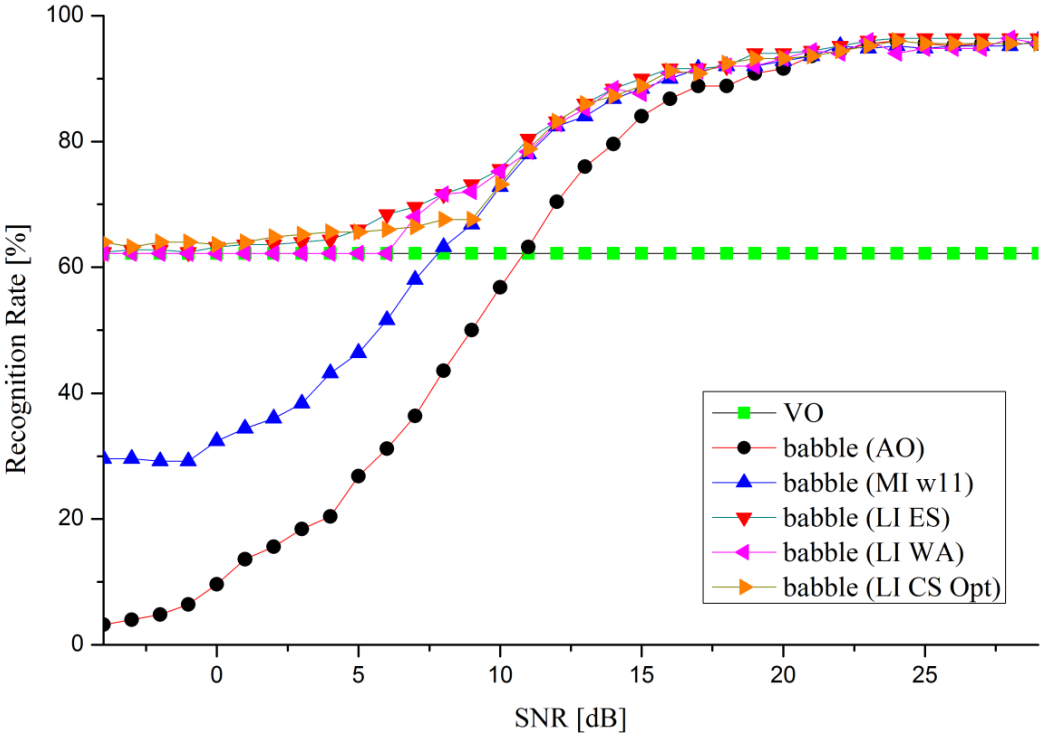


Fig.8.20: Comparison of fusion strategies in babble noise.

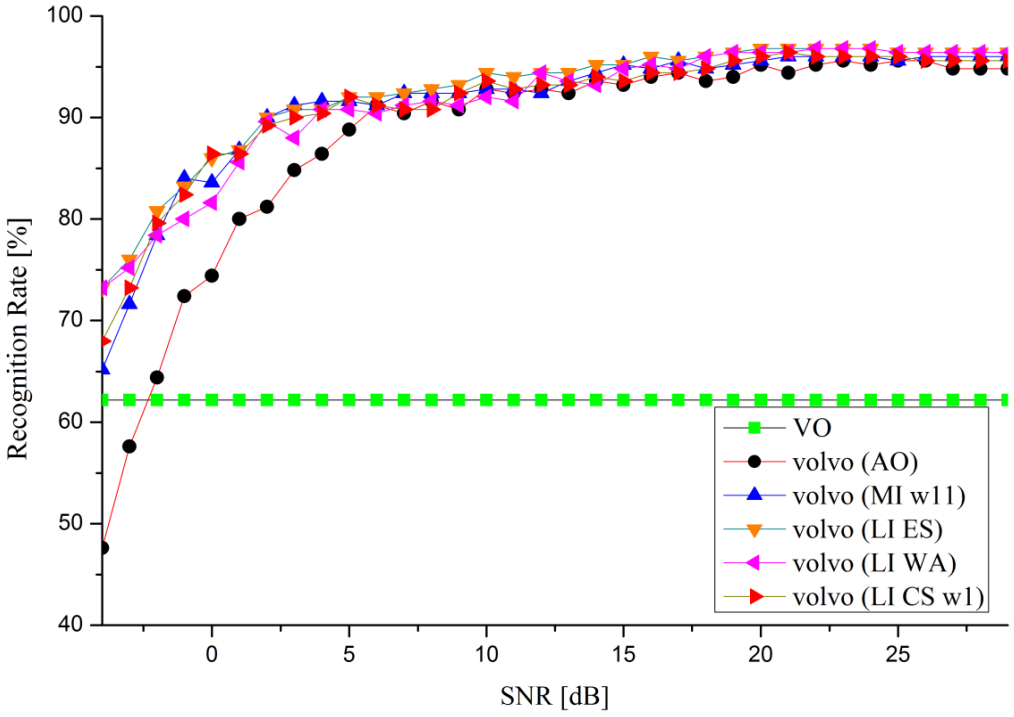


Fig.8.21: Comparison of fusion strategies in volvo noise.

Chapter 9

Conclusion

9.1 Thesis achievements

In this work, the following tasks have been proposed, described and experimentally evaluated in order to get a practically applicable ASR system for Vietnamese:

1. An efficient way to collect large text and speech corpora from publically available sources.
2. Design of an optimized phoneme set and a grapheme-to-phoneme transducer.
3. Design of a novel method to build syllable and multi-syllable pronunciation dictionary with high (99%) coverage of common Vietnamese texts.
4. Design of a language model that takes into account several tens of thousands frequently collocated syllables fits them into a bigram LM.
5. A comparative study to find the optimal way of incorporating tone into ASR systems for Vietnamese.
6. A large experimental evaluation of different approaches to build an LVCSR for Vietnamese, which resulted in a system that yields syllable accuracy close to 80% on broadcast speech.
7. A study focused on audio-visual speech recognition task.

The above mentioned achievements are discussed in more details in the following text.

9.2 Text and speech corpora

For experiments on speech recognition tasks, three types of text and speech corpora have been collected in this work:

First, the text corpora were extracted from the Internet's resource which contained text of several categories such as news, literature, etc. Both the resulted general and specific purpose text corpora were used in the tasks of LM evaluation and LVCSR. The total text corpus contains more than 8 million sentences with about 180 million syllables.

Second, the speech corpus for experiments on LVCSR was also collected from the Internet. This corpus contains speech of several categories including story reading, news report, weather forecast, conversation, etc., and covers three main dialects of Vietnamese. The total of 24871 utterances were selected from the audio files and manually transcribed to obtain the final speech corpus with the length of 50 hours 22 minutes. The number of speaker in this corpus is 196 (69 male speakers and 127 female speakers).

Finally, the audio-visual speech corpus was constructed for isolated word speech recognition task. This corpus was recorded from 50 speakers in room condition and contained two data sets. The first set contained continuous speech of 2500 utterances of 50 adaptation sentences and another 2500 utterances of 2500 specific sentences where each speaker was asked to utter 100 sentences. The second set consisted of speech data of 50 isolated words which were recorded from each of the 50 speakers.

9.3 Tone hypotheses

This thesis has dealt with the most difficult and also the most interesting problem in LVCSR of Vietnamese: modeling tone in syllable. By solving three main hypotheses about tone, the author provided not only the insightful information about properties of tone in syllable but also the baseline method to integrate tone into recognizer of LVCSR tasks.

In the first hypothesis, which related to whether or not tone should be modeled in syllable, the recognition accuracies in Tab. 7.16 showed that for all of speech recognition strategies as well as phoneme set types, methods with tonal models usually outperform methods without tone modeling. And so, it is true to say that tone is an important component of a Vietnamese syllable and has to be modeled one or another way to obtain optimized results for LVCSR tasks.

Reaffirmation of the first hypothesis, the second hypothesis dealt with the problem of what is the role of tone in syllables. The experiments showed that for the same analyzing strategy, methods based on dependent tone always give better results than methods with independent tone. This leads to the fact that independent tone cannot be properly modeled in the same manner with other phonetic units in a syllable. Based on this hypothesis, two methods have been proposed to solve the tonal role problem. The first method uses the so-called multi-stream HMMs to model tone completely independent with other phonetic unit types. Although this method has its own obstacles, interesting results can be obtained when

more works are put on it. For the second method, which was examined in this thesis, tone is modeled independently in the same manner with other phonetic units, but by creating the relation between tone and other phonetic units in the form of context-dependent HMMs, tone is better modeled and the recognizer can better deal with tone.

For the final hypothesis, the task of examining the position of tone in Vietnamese syllable is solved. It could be seen that in the same speech recognition strategy as well as phoneme set type, the methods where tone is located at the end of syllable give better result than the methods where tone is located after main vowel both with dependent and independent tone hypotheses. This means the important part of tone is located at the end of Vietnamese syllable and should be emphasized in recognizers of LVCSR tasks.

9.4 Audio speech recognition of Vietnamese

The first contribution of this work was the proposal of a standard phoneme set which was the core of all experiments on ASR of Vietnamese. This phoneme set along with the grapheme-to-phoneme mapping table will make the researches on ASR of Vietnamese more understandable. Initial work based on the above material has been presented in [P1].

To deal with ASR tasks, four different strategies for speech recognition of Vietnamese were examined. In these strategies, syllable-based methods are more suitable for isolated Vietnamese syllable or isolated word tasks which provide very high recognition rate. Experiments in this work showed that recognition rate of 97% could be achieved for the task of recognition of 50 isolated words. The other three strategies including phoneme-based, vowel-based and rhyme-based methods were applied to the LVCSR task with different degree of successes. The vowel-based and rhyme-based strategies tended to give better recognition rate in context-independent LVCSR task in which the rhyme-based strategy obtained the highest accuracy using method *C1RT_D* (65.26%). On the other hand, the phoneme-based strategy was more flexible and provided better results in context-dependent LVCSR task. The method *C1wVC2T_I* has proved to be the most appropriate method for dealing with LVCSR of Vietnamese which not only fit the hypotheses about tone but also suitable for context-dependent HMMs strategy. It could obtain the recognition rate of 73.90% in comparison with 65.26% of the method *C1RT_D* on the same training and testing data.

Also, in this work, the effects of different types of LM were examined. First, the results showed that LM estimated from the general purpose text corpus was good enough for LVCSR task. This LM resulted in the syllable accuracy of 72.31% in comparison with 73.90% of LM estimated from the combined text corpus. It also showed that LMs estimated using Kneser-Ney with interpolating model or backoff model as smoothing methods gave the best results for speech recognition of Vietnamese. An interesting aspect of syllable-based LM is that LMs with vocabulary size of 6000 to 7000 syllables are feasible for LVCSR task. Initial experiments on word-based LM in the form of multi-syllable-based LM also gave some promising results. The best syllable accuracy was 78.70% obtained with the vocabulary of 40000 bi-syllabic tokens. It proved that by enlarging the text corpus and improving the word segmentation algorithms, speech recognizers based on word-based LM can obtain very good results.

To further improve the recognition rate, a gender dependent recognizer was applied to the LVCSR task. This optimization strategy has shown some improvements in syllable accuracy in which the highest result of 79.66% was obtained with multi-syllable-based LM using Kneser-Ney smoothing.

Also, to make it more interesting, I will make some references between this work with one of the most interesting paper on LVCSR of Vietnamese at current time which is presented by Vu and Schultz in [125]. There are several differences between the two works in which the work in [125] provides more standard data and optimized strategies. First, they used a standard speech corpus where speech data was recorded with a headset microphone in clean environmental conditions. The text for speech corpus and LM construction is of the same category. On the other hand, this work used a rather challenging speech corpus where speech data was collected from the Internet resource with different qualities, noise conditions, recorded devices, dialects, etc. Also, the text corpus in this work is tended for general purpose. Second, they used a more optimized LM and accoustic model for speech recognition task in which a tri-gram LM was evaluated in comparison with bi-gram LM in this work. Also, they used a feature vector with more dimension and applied linear discriminant analysis to improve the feature vector. In this work a rather standard feature vector was used instead. When training HMMs, they used the selected MM7 models as seed models to produce initial state alignments, this work used flat-start procedure. In the

summary, the work in [125] is more concentrated in the optimization of the system, while in this work, I am more concentrated in the methodology, strategy to deal with tone as well as speech recognition tasks. This makes the work in this thesis is more suitable for research purpose and, comparing to their baseline recognizer, our recognizers gives comparable result. And so, it is believed that, when the same optimized strategies is applied for the best method in this work, a really good result can be obtained.

9.5 Audio-visual speech recognition

In this work, two sets of experiments on visual speech analysis including experiments on visual only and experiments on audio-visual isolated word speech recognition of Vietnamese were examined.

For visual only isolated word speech recognition task, the performance of two visual front ends for feature extraction was first compared. The recognition results showed that the HLDA visual front end outperformed the 1-Stage LDA visual front end in both the highest accuracy (62%) and average accuracy, and both visual front ends did improve the static visual feature. Then using the HLDA visual front end for feature extraction, three different types of visual feature including DCT, PCA and AAM were studied. The best visual only recognition results were obtained with DCT and PCA feature types (62%), but DCT feature provided better dimensional reduction (16 coefficients). It also showed that, DCT and PCA feature outperformed the AAM feature when using HLDA as the basic visual front end.

For audio-visual isolated word speech recognition task, two different fusion strategies including middle integration and late integration were studied. First it is worth to mention that different types of noise affect the audio stream in different degree, and in high noise condition the performance of audio only recognizer degrade rapidly. Using middle integration strategy, the results showed that audio-visual recognizer outperformed audio only recognizer especially in high noise condition and middle integration strategy using adaptation data outperformed the case where both audio and visual stream weights equal to 1. In late integration strategy, method using exhausted search outperformed the audio only and visual only in most of the noise condition. The results also showed that using adapting data for audio weigh searching can obtain the results as good as the optimal case. On the other hand, method using automatic weight selection obtained its optimal results depended

on many factors including the number of N-best hypotheses, the weight adjusting coefficient and how the confidence score is computed.

In summary, the fusion of audio and visual stream using middle integration or late integration outperformed the audio only speech recognition task in which the late integration performed better in many aspects such as recognition rate, runtime stream weight modification, the independence in the combination of two streams output, etc., and so it makes late integration more applicable than other fusion strategies.

9.6 Future work

The extension to this work can be done in many aspects which tend to be crucial not only for the improvement in performance of ASR tasks but also for the development of ASR of Vietnamese.

First, more studies on tone, especially properties of tone in continuous speech are key for the total understanding of tone in natural speech of Vietnamese. Several researches have been done to deal with tone in the isolated word case but there is not any works in the continuous speech case making tone still a big challenge for any researchers that want to work with Vietnamese.

For experiment on LVCSR, only gender dependent optimization method is applied in this work. It is believed that the recognition accuracy can be improved dramatically when applying other optimization methods on two aspects: Vietnamese specific optimization strategies (using word-based LM, extracting pitch information, more tone processing methods, dealing with dialect, etc.) and general optimization strategies (signal adaptation, dealing with noise, LM tuning, etc.).

Finally, experiments on AVSR show that visual stream can vastly improve ASR of Vietnamese, especially in noise condition. But, the main part of the experiments was to examine the effect of visual features on the isolated word speech recognition task. Many studies in this field have obtained good results also in the continuous speech recognition task. So, it would be required to extend AVSR to the task of continuous speech.

Appendix A

Basic Phonetic Unit Set

A.1. Phoneme set of phoneme-based strategy

Tab. A.1: Phone-based phoneme set: C1wVC2

aa	ah	ao	au	ax	b	d	ee
eh	f	g	h	ch	ie	ih	iy
k	kh	l	m	n	ng	nh	oo
p	r	s	sh	t	th	tr	ua
uh	uo	uw	v	w	y	zh	

Tab. A.2: Grapheme-based phoneme set: C1wVC2

a	a1	a2	b	c	d	d1	e
e1	g	gh	gi	h	ch	i	ia
ie1	k	kh	l	m	n	ng	ngh
nh	o	o1	o2	oo	p	ph	q
r	s	t	th	tr	u	u1	u1a
u1o2	ua	uo1	v	x	y	ya	ye1

Tab. A.3: Phone-based phoneme set: C1wVTC2T_I, C1wVC2T_I, C1wVTC2_I

aa	ah	ao	au	ax	b	d	ee	eh
f	g	h	ch	ie	ih	iy	k	kh
l	m	n	ng	nh	oo	p	r	s
sh	t	th	tr	ua	uh	uo	uw	v
w	y	z1	z2	z3	z4	z5	z6	zh

Tab. A.4: Grapheme-based phoneme set: C1wVTC2T_I, C1wVC2T_I, C1wVTC2_I

a	a1	a2	b	c	d	d1	e	e1
g	gh	gi	h	ch	i	ia	ie1	k
kh	l	m	n	ng	ngh	nh	o	o1
o2	oo	p	ph	q	r	s	t	th
tr	u	u1	u1a	u1o2	ua	uo1	v	x
y	ya	ye1	z1	z2	z3	z4	z5	z6

Tab. A.5: Phone-based phoneme set: C1wVC2T_D

aa	aaz1	aaz2	aaz3	aaz4	aaz5	aaz6	ah
ao	aoz1	aoz2	aoz3	aoz4	aoz5	aoz6	au
ax	axz1	axz2	axz3	axz4	axz5	axz6	b
d	ee	eez1	eez2	eez3	eez4	eez5	eez6
eh	ehz1	ehz2	ehz3	ehz4	ehz5	ehz6	f
g	h	ch	chz5	chz6	ie	iez1	iez2
iez3	iez4	iez5	iez6	ihz1	ihz2	ihz3	ihz4
ihz5	ihz6	iy	iyz1	iyz2	iyz3	iyz4	iyz5
iyz6	k	kh	kz5	kz6	l	m	mz1
mz2	mz3	mz4	mz5	mz6	n	ng	ngz1
ngz2	ngz3	ngz4	ngz5	ngz6	nh	nhz1	nhz2
nhz3	nhz4	nhz5	nhz6	nz1	nz2	nz3	nz4
nz5	nz6	oo	ooz1	ooz2	ooz3	ooz4	ooz5
ooz6	pz5	pz6	r	s	sh	t	th
tr	tz5	tz6	ua	uaz1	uaz2	uaz3	uaz4
uaz5	uaz6	uh	uhz1	uhz2	uhz3	uhz4	uhz5
uhz6	uo	uoz1	uoz2	uoz3	uoz4	uoz5	uoz6
uw	uwz1	uwz2	uwz3	uwz4	uwz5	uwz6	v
w	wz1	wz2	wz3	wz4	wz5	wz6	y
zh							

Tab. A.6: Grapheme-based phoneme set: C1wVC2T_D

a	a1	a2	az1	az2	az3	az4	az5
az6	b	c	cz5	cz6	d	d1	e
e1	e1z1	e1z2	e1z3	e1z4	e1z5	e1z6	ez1
ez2	ez3	ez4	ez5	ez6	g	gh	gi
h	ch	chz5	chz6	i	iaz1	iaz2	iaz3
iaz4	iaz5	iaz6	ie1	iz1	iz2	iz3	iz4
iz5	iz6	k	kh	l	m	mz1	mz2
mz3	mz4	mz5	mz6	n	ng	ngh	ngz1
ngz2	ngz3	ngz4	ngz5	ngz6	nh	nhz1	nhz2
nhz3	nhz4	nhz5	nhz6	nz1	nz2	nz3	nz4
nz5	nz6	o	o1	o1z1	o1z2	o1z3	o1z4
o1z5	o1z6	o2	o2z1	o2z2	o2z3	o2z4	o2z5
o2z6	oo	oz1	oz2	oz3	oz4	oz5	oz6
ph	pz5	pz6	q	r	s	t	th
tr	tz5	tz6	u	u1	u1az1	u1az2	u1az3
u1az4	u1az5	u1az6	u1o2	u1z1	u1z2	u1z3	u1z4
u1z5	u1z6	uaz1	uaz2	uaz3	uaz4	uaz5	uaz6
uo1	uz1	uz2	uz3	uz4	uz5	uz6	v
x	y	yaz1	ye1	yz1	yz2	yz3	yz4
yz5	yz6						

Tab. A.7: Phone-based phoneme set: C1wVTC2_D

aaz1	aaz2	aaz3	aaz4	aaz5	aaz6	ahz1
ahz2	ahz3	ahz4	ahz5	ahz6	aoz1	aoz2
aoz3	aoz4	aoz5	aoz6	auz1	auz2	auz3
auz4	auz5	auz6	axz1	axz2	axz3	axz4
axz5	axz6	b	d	eez1	eez2	eez3
eez4	eez5	eez6	ehz1	ehz2	ehz3	ehz4
ehz5	ehz6	f	g	h	ch	iez1
iez2	iez3	iez4	iez5	iez6	ih	iyz1
iyz2	iyz3	iyz4	iyz5	iyz6	k	kh
l	m	n	ng	nh	ooz1	ooz2
ooz3	ooz4	ooz5	ooz6	p	r	s
sh	t	th	tr	uaz1	uaz2	uaz3
uaz4	uaz5	uaz6	uhz1	uhz2	uhz3	uhz4
uhz5	uhz6	uoz1	uoz2	uoz3	uoz4	uoz5
uoz6	uwz1	uwz2	uwz3	uwz4	uwz5	uwz6
v	w	y	zh			

Tab. A.8: Grapheme-based phoneme set: C1wVTC2_D

a1z1	a1z2	a1z3	a1z4	a1z5	a1z6	a2z1
a2z2	a2z3	a2z4	a2z5	a2z6	az1	az2
az3	az4	az5	az6	b	c	d
d1	e1z1	e1z2	e1z3	e1z4	e1z5	e1z6
ez1	ez2	ez3	ez4	ez5	ez6	g
gh	gi	h	ch	i	iaz1	iaz2
iaz3	iaz4	iaz5	iaz6	ie1z1	ie1z2	ie1z3
ie1z4	ie1z5	ie1z6	iz1	iz2	iz3	iz4
iz5	iz6	k	kh	l	m	n
ng	ngh	nh	o	o1z1	o1z2	o1z3
o1z4	o1z5	o1z6	o2z1	o2z2	o2z3	o2z4
o2z5	o2z6	ooz1	ooz2	ooz5	oz1	oz2
oz3	oz4	oz5	oz6	p	ph	q
r	s	t	th	tr	u	u1az1
u1az2	u1az3	u1az4	u1az5	u1az6	u1o2z1	u1o2z2
u1o2z3	u1o2z4	u1o2z5	u1o2z6	u1z1	u1z2	u1z3
u1z4	u1z5	u1z6	uaz1	uaz2	uaz3	uaz4
uaz5	uaz6	uo1z1	uo1z2	uo1z3	uo1z4	uo1z5
uo1z6	uz1	uz2	uz3	uz4	uz5	uz6
v	x	y	yaz1	ye1z1	ye1z2	ye1z3
ye1z4	ye1z5	ye1z6	yz1	yz2	yz3	yz4
yz5	yz6					

Tab. A.9: Phone-based phoneme set: C1wVTC2T_D

aaz1	aaz2	aaz3	aaz4	aaz5	aaz6	ahz1	ahz2	ahz3	ahz4	ahz5	ahz6
aoz1	aoz2	aoz3	aoz4	aoz5	aoz6	auz1	auz2	auz3	auz4	auz5	auz6
axz1	axz2	axz3	axz4	axz5	axz6	b	d	eez1	eez2	eez3	eez4
eez5	eez6	ehz1	ehz2	ehz3	ehz4	ehz5	ehz6	f	g	h	ch
chz5	chz6	iez1	iez2	iez3	iez4	iez5	iez6	ihz1	ihz2	ihz3	ihz4
ihz5	ihz6	iyz1	iyz2	iyz3	iyz4	iyz5	iyz6	k	kh	kz5	kz6
l	m	mz1	mz2	mz3	mz4	mz5	mz6	n	ng	ngz1	ngz2
ngz3	ngz4	ngz5	ngz6	nh	nhz1	nhz2	nhz3	nhz4	nhz5	nhz6	nz1
nz2	nz3	nz4	nz5	nz6	ooz1	ooz2	ooz3	ooz4	ooz5	ooz6	pz5
pz6	r	s	sh	t	th	tr	tz5	tz6	uaz1	uaz2	uaz3
uaz4	uaz5	uaz6	uhz1	uhz2	uhz3	uhz4	uhz5	uhz6	uoz1	uoz2	uoz3
uoz4	uoz5	uoz6	uwz1	uwz2	uwz3	uwz4	uwz5	uwz6	v	w	wz1
wz2	wz3	wz4	wz5	wz6	y	zh					

Tab. A.10: Grapheme-based phoneme set: C1wVTC2T_D

a1z1	a1z2	a1z3	a1z4	a1z5	a1z6	a2z1	a2z2	a2z3	a2z4	a2z5
a2z6	az1	az2	az3	az4	az5	az6	b	c	cz5	cz6
d	d1	e1z1	e1z2	e1z3	e1z4	e1z5	e1z6	ez1	ez2	ez3
ez4	ez5	ez6	g	gh	gi	h	ch	chz5	chz6	iaz1
iaz2	iaz3	iaz4	iaz5	iaz6	ie1z1	ie1z2	ie1z3	ie1z4	ie1z5	ie1z6
iz1	iz2	iz3	iz4	iz5	iz6	k	kh	l	m	mz1
mz2	mz3	mz4	mz5	mz6	n	ng	ngh	ngz1	ngz2	ngz3
ngz4	ngz5	ngz6	nh	nhz1	nhz2	nhz3	nhz4	nhz5	nhz6	nz1
nz2	nz3	nz4	nz5	nz6	o	o1z1	o1z2	o1z3	o1z4	o1z5
o1z6	o2z1	o2z2	o2z3	o2z4	o2z5	o2z6	ooz1	ooz2	ooz5	oz1
oz2	oz3	oz4	oz5	oz6	ph	pz5	pz6	q	r	s
t	th	tr	tz5	tz6	u	u1az1	u1az2	u1az3	u1az4	u1az5
u1az6	u1o2z1	u1o2z2	u1o2z3	u1o2z4	u1o2z5	u1o2z6	u1z1	u1z2	u1z3	u1z4
u1z5	u1z6	uaz1	uaz2	uaz3	uaz4	uaz5	uaz6	uo1z1	uo1z2	uo1z3
uo1z4	uo1z5	uo1z6	uz1	uz2	uz3	uz4	uz5	uz6	v	x
yaz1	ye1z1	ye1z2	ye1z3	ye1z4	ye1z5	ye1z6	yz1	yz2	yz3	yz4
yz5	yz6									

A.2. Phoneme set of vowel-based strategy

Tab. A.11: Phone-based phoneme set: C1MC

aa	aaih	aaw	ah	ahih	ahw	ao	aoih
au	auih	auw	ax	axih	b	d	ee
eew	eh	ehw	f	g	h	ch	ie
iew	iy	iyw	k	kh	l	m	n
ng	nh	oo	ooih	p	r	s	sh
t	th	tr	ua	uaih	uaw	uh	uhih
uhw	uo	uoih	uw	v	waa	waaih	waaw
wah	wahih	wao	wau	wauih	wauw	wax	waxih
wee	weh	wehw	wie	wiy	wiyw	y	zh

Tab. A.12: Grapheme-based phoneme set: C1MC

a	a1	a2	a2u	a2y	ai	ao	au
ay	b	c	d	d1	e	e1	e1u
eo	g	gh	gi	h	ch	i	ia
ie1	ie1u	iu	k	kh	l	m	n
ng	ngh	nh	o	o1	o1i	o2	o2i
oa	oa1	oai	oao	oay	oe	oeo	oi
oo	p	ph	q	r	s	t	th
tr	u	u1	u1a	u1i	u1o2	u1o2i	u1o2u
u1u	ua	ua1	ua2	ua2y	uai	uao	uau
uay	ue	ue1	ueo	ui	uo1	uo1i	uo2
uo2i	uy	uya	uye1	uyu	v	x	y
ye1	ye1u						

Tab. A.13: Phone-based phoneme set: C1MCT_I, C1MTC_I

aa	aaih	aaw	ah	ahih	ahw	ao	aoih
au	auih	auw	ax	axih	b	d	ee
eew	eh	ehw	f	g	h	ch	ie
iew	iy	iyw	k	kh	l	m	n
ng	nh	oo	ooih	p	r	s	sh
t	th	tr	ua	uaih	uaw	uh	uhih
uhw	uo	uoih	uw	v	waa	waaih	waaw
wah	wahih	wao	wau	wauih	wauw	wax	waxih
wee	weh	wehw	wie	wiy	wiyw	y	z1
z2	z3	z4	z5	z6	zh		

Tab. A.14: Grapheme-based phoneme set: C1MCT_I, C1MTC_I

a	a1	a2	a2u	a2y	ai	ao	au
ay	b	c	d	d1	e	e1	e1u
eo	g	gh	gi	h	ch	i	ia
ie1	ie1u	iu	k	kh	l	m	n
ng	ngh	nh	o	o1	o1i	o2	o2i
oa	oa1	oai	oao	oay	oe	oeo	oi
oo	p	ph	q	r	s	t	th
tr	u	u1	u1a	u1i	u1o2	u1o2i	u1o2u
u1u	ua	ua1	ua2	ua2y	uai	uao	uau
uay	ue	ue1	ueo	ui	uo1	uo1i	uo2
uo2i	uy	uya	uye1	uyu	v	x	y
ye1	ye1u	z1	z2	z3	z4	z5	z6

Tab. A.15: Phone-based phoneme set: C1MCT_D

aa	aaihz1	aaihz2	aaihz3	aaihz4	aaihz5	aaihz6	aawz1
aawz2	aawz3	aawz4	aawz5	aawz6	aaz1	aaz2	aaz3
aaz4	aaz5	aaz6	ah	ahihz1	ahihz2	ahihz3	ahihz4
ahihz5	ahihz6	ahwz1	ahwz2	ahwz3	ahwz4	ahwz5	ahwz6
ao	aoihz1	aoihz2	aoihz3	aoihz4	aoihz5	aoihz6	aoz1
aoz2	aoz3	aoz4	aoz5	aoz6	au	auihz1	auihz2
auihz3	auihz4	auihz5	auihz6	auwz1	auwz2	auwz4	auwz5
auwz6	ax	axihz1	axihz2	axihz3	axihz4	axihz5	axihz6
axz1	axz2	axz3	axz4	axz5	axz6	b	d
ee	eewz1	eewz2	eewz3	eewz4	eewz5	eewz6	eez1
eez2	eez3	eez4	eez5	eez6	eh	ehwz1	ehwz2
ehwz3	ehwz4	ehwz5	ehwz6	ehz1	ehz2	ehz3	ehz4
ehz5	ehz6	f	g	h	ch	chz5	chz6
ie	iewz1	iewz2	iewz3	iewz4	iewz5	iewz6	iez1
iez2	iez3	iez4	iez5	iez6	iy	iywz1	iywz2
iywz3	iywz4	iywz5	iywz6	iyz1	iyz2	iyz3	iyz4
iyz5	iyz6	k	kh	kz5	kz6	l	m
mz1	mz2	mz3	mz4	mz5	mz6	n	ng
ngz1	ngz2	ngz3	ngz4	ngz5	ngz6	nh	nhz1
nhz2	nhz3	nhz4	nhz5	nhz6	nz1	nz2	nz3
nz4	nz5	nz6	oo	ooihz1	ooihz2	ooihz3	ooihz4
ooihz5	ooihz6	ooz1	ooz2	ooz3	ooz4	ooz5	ooz6
pz5	pz6	r	s	sh	t	th	tr
tz5	tz6	ua	uaihz1	uaihz2	uaihz3	uaihz4	uaihz5
uaihz6	uawz1	uawz5	uawz6	uaz1	uaz2	uaz3	uaz4
uaz5	uaz6	uh	uhihz4	uhwz1	uhwz2	uhwz3	uhwz4
uhwz5	uhwz6	uhz1	uhz2	uhz3	uhz4	uhz5	uhz6
uo	uoihz1	uoihz2	uoihz3	uoihz4	uoihz5	uoihz6	uoz1
uoz2	uoz3	uoz4	uoz5	uoz6	uw	uwz1	uwz2
uwz3	uwz4	uwz5	uwz6	v	waa	waaihz1	waaihz2
waaihz3	waaihz4	waaihz5	waaihz6	waawz5	waawz6	waaz1	waaz2
waaz3	waaz4	waaz5	waaz6	wah	wahihz1	wahihz2	wahihz3
wahihz4	wahihz5	wahihz6	wao	wau	wauihz1	wauihz2	wauihz4
wauihz5	wauihz6	wauwz6	wax	waxihz5	waxz1	waxz2	waxz4
wee	weez1	weez2	weez4	weez5	weez6	weh	wehwz1
wehwz2	wehwz4	wehwz5	wehwz6	wehz1	wehz2	wehz3	wehz4
wehz5	wehz6	wie	wiez1	wiy	wiywz4	wiywz6	wiyz1
wiyz2	wiyz3	wiyz4	wiyz5	wiyz6	y	zh	

Tab. A.16: Grapheme-based phoneme set: C1MCT_D

a	a1	a2	a2uz1	a2uz2	a2uz3	a2uz4	a2uz5
a2uz6	a2yz1	a2yz2	a2yz3	a2yz4	a2yz5	a2yz6	aiz1
aiz2	aiz3	aiz4	aiz5	aiz6	aoz1	aoz2	aoz3
aoz4	aoz5	aoz6	auz1	auz2	auz4	auz5	auz6
ayz1	ayz2	ayz3	ayz4	ayz5	ayz6	az1	az2
az3	az4	az5	az6	b	c	cz5	cz6
d	d1	e	e1	e1uz1	e1uz2	e1uz3	e1uz4
e1uz5	e1uz6	e1z1	e1z2	e1z3	e1z4	e1z5	e1z6
eoz1	eoz2	eoz3	eoz4	eoz5	eoz6	ez1	ez2
ez3	ez4	ez5	ez6	g	gh	gi	h
ch	chz5	chz6	i	iaz1	iaz2	iaz3	iaz4
iaz5	iaz6	ie1	ie1uz1	ie1uz2	ie1uz3	ie1uz4	ie1uz5
ie1uz6	iuz1	iuz2	iuz3	iuz4	iuz5	iuz6	iz1
iz2	iz3	iz4	iz5	iz6	k	kh	l
m	mz1	mz2	mz3	mz4	mz5	mz6	n
ng	ngh	ngz1	ngz2	ngz3	ngz4	ngz5	ngz6
nh	nhz1	nhz2	nhz3	nhz4	nhz5	nhz6	nz1
nz2	nz3	nz4	nz5	nz6	o	o1	o1iz1
o1iz2	o1iz3	o1iz4	o1iz5	o1iz6	o1z1	o1z2	o1z3
o1z4	o1z5	o1z6	o2	o2iz1	o2iz2	o2iz3	o2iz4
o2iz5	o2iz6	o2z1	o2z2	o2z3	o2z4	o2z5	o2z6
oa	oa1	oaiz1	oaiz2	oaiz3	oaiz4	oaiz5	oaiz6
oaoz5	oayz1	oayz4	oayz5	oaz1	oaz2	oaz3	oaz4
oaz5	oaz6	oe	oeoz1	oeoz2	oeoz4	oeoz5	oeoz6
oez1	oez2	oez3	oez4	oez5	oez6	oiz1	oiz2
oiz3	oiz4	oiz5	oiz6	oo	oz1	oz2	oz3
oz4	oz5	oz6	ph	pz5	pz6	q	r
s	t	th	tr	tz5	tz6	u	u1
u1az1	u1az2	u1az3	u1az4	u1az5	u1az6	u1iz4	u1o2
u1o2iz1	u1o2iz2	u1o2iz3	u1o2iz4	u1o2iz5	u1o2iz6	u1o2uz1	u1o2uz5
u1o2uz6	u1uz1	u1uz2	u1uz3	u1uz4	u1uz5	u1uz6	u1z1
u1z2	u1z3	u1z4	u1z5	u1z6	ua	ua1	ua2
ua2yz1	ua2yz2	ua2yz3	ua2yz4	ua2yz5	ua2yz6	uaiz1	uaiz2
uaiz5	uaiz6	uaoz5	uaoz6	uauz6	uayz1	uayz2	uayz4
uayz6	uaz1	uaz2	uaz3	uaz4	uaz5	uaz6	ue
ue1	ue1z1	ue1z2	ue1z4	ue1z5	ue1z6	ueoz1	ueoz5
ueoz6	uez1	uez2	uez3	uez4	ui	uiz1	uiz2
uiz3	uiz4	uiz5	uiz6	uo1	uo1iz1	uo1iz2	uo1iz3
uo1iz4	uo1iz5	uo1iz6	uo2	uo2iz5	uo2z1	uo2z2	uo2z4
uy	uyaz1	uye1	uyuz4	uyuz6	uyz1	uyz2	uyz3
uyz4	uyz5	uyz6	uz1	uz2	uz3	uz4	uz5
uz6	v	x	ye1	ye1uz1	ye1uz4	ye1uz5	yz1
yz2	yz3	yz4	yz5	yz6			

Tab. A.17: Phone-based phoneme set: C1MTC_D

aaihz1	aaihz2	aaihz3	aaihz4	aaihz5	aaihz6	aawz1	aawz2
aawz3	aawz4	aawz5	aawz6	aaz1	aaz2	aaz3	aaz4
aaz5	aaz6	ahihz1	ahihz2	ahihz3	ahihz4	ahihz5	ahihz6
ahwz1	ahwz2	ahwz3	ahwz4	ahwz5	ahwz6	ahz1	ahz2
ahz3	ahz4	ahz5	ahz6	aoihz1	aoihz2	aoihz3	aoihz4
aoihz5	aoihz6	aoz1	aoz2	aoz3	aoz4	aoz5	aoz6
auihz1	auihz2	auihz3	auihz4	auihz5	auihz6	auwz1	auwz2
auwz4	auwz5	auwz6	auz1	auz2	auz3	auz4	auz5
auz6	axihz1	axihz2	axihz3	axihz4	axihz5	axihz6	axz1
axz2	axz3	axz4	axz5	axz6	b	d	eewz1
eewz2	eewz3	eewz4	eewz5	eewz6	eez1	eez2	eez3
eez4	eez5	eez6	ehwz1	ehwz2	ehwz3	ehwz4	ehwz5
ehwz6	ehz1	ehz2	ehz3	ehz4	ehz5	ehz6	f
g	h	ch	iewz1	iewz2	iewz3	iewz4	iewz5
iewz6	iez1	iez2	iez3	iez4	iez5	iez6	iywz1
iywz2	iywz3	iywz4	iywz5	iywz6	iyz1	iyz2	iyz3
iyz4	iyz5	iyz6	k	kh	l	m	n
ng	nh	ooihz1	ooihz2	ooihz3	ooihz4	ooihz5	ooihz6
ooz1	ooz2	ooz3	ooz4	ooz5	ooz6	p	r
s	sh	t	th	tr	uaihz1	uaihz2	uaihz3
uaihz4	uaihz5	uaihz6	uawz1	uawz5	uawz6	uaz1	uaz2
uaz3	uaz4	uaz5	uaz6	uhihz4	uhwz1	uhwz2	uhwz3
uhwz4	uhwz5	uhwz6	uhz1	uhz2	uhz3	uhz4	uhz5
uhz6	uoihz1	uoihz2	uoihz3	uoihz4	uoihz5	uoihz6	uoz1
uoz2	uoz3	uoz4	uoz5	uoz6	uwz1	uwz2	uwz3
uwz4	uwz5	uwz6	v	waaihz1	waaihz2	waaihz3	waaihz4
waaihz5	waaihz6	waawz5	waawz6	waaz1	waaz2	waaz3	waaz4
waaz5	waaz6	wahihz1	wahihz2	wahihz3	wahihz4	wahihz5	wahihz6
wahz1	wahz2	wahz3	wahz4	wahz5	wahz6	waoz5	wauihz1
wauihz2	wauihz4	wauihz5	wauihz6	wauwz6	wauz1	wauz2	wauz3
wauz4	wauz5	wauz6	waxihz5	waxz1	waxz2	waxz3	waxz4
waxz5	weez1	weez2	weez4	weez5	weez6	wehwz1	wehwz2
wehwz4	wehwz5	wehwz6	wehz1	wehz2	wehz3	wehz4	wehz5
wehz6	wiez1	wiez2	wiez3	wiez4	wiez5	wiez6	wiywz4
wiywz6	wiyz1	wiyz2	wiyz3	wiyz4	wiyz5	wiyz6	y
zh							

Tab. A.18: Grapheme-based phoneme set: C1MTC_D

a1z1	a1z2	a1z3	a1z4	a1z5	a1z6	a2uz1	a2uz2
a2uz3	a2uz4	a2uz5	a2uz6	a2yz1	a2yz2	a2yz3	a2yz4
a2yz5	a2yz6	a2z1	a2z2	a2z3	a2z4	a2z5	a2z6
aiz1	aiz2	aiz3	aiz4	aiz5	aiz6	aoz1	aoz2
aoz3	aoz4	aoz5	aoz6	auz1	auz2	auz4	auz5
auz6	ayz1	ayz2	ayz3	ayz4	ayz5	ayz6	az1
az2	az3	az4	az5	az6	b	c	d
d1	e1uz1	e1uz2	e1uz3	e1uz4	e1uz5	e1uz6	e1z1
e1z2	e1z3	e1z4	e1z5	e1z6	eoz1	eoz2	eoz3
eoz4	eoz5	eoz6	ez1	ez2	ez3	ez4	ez5
ez6	g	gh	gi	h	ch	iaz1	iaz2
iaz3	iaz4	iaz5	iaz6	ie1uz1	ie1uz2	ie1uz3	ie1uz4
ie1uz5	ie1uz6	ie1z1	ie1z2	ie1z3	ie1z4	ie1z5	ie1z6
iuz1	iuz2	iuz3	iuz4	iuz5	iuz6	iz1	iz2
iz3	iz4	iz5	iz6	k	kh	l	m
n	ng	ngh	nh	o1iz1	o1iz2	o1iz3	o1iz4
o1iz5	o1iz6	o1z1	o1z2	o1z3	o1z4	o1z5	o1z6
o2iz1	o2iz2	o2iz3	o2iz4	o2iz5	o2iz6	o2z1	o2z2
o2z3	o2z4	o2z5	o2z6	oa1z1	oa1z2	oa1z3	oa1z4
oa1z5	oa1z6	oaiz1	oaiz2	oaiz3	oaiz4	oaiz5	oaiz6
oaoz5	oayz1	oayz4	oayz5	oaz1	oaz2	oaz3	oaz4
oaz5	oaz6	oeoz1	oeoz2	oeoz4	oeoz5	oeoz6	oez1
oez2	oez3	oez4	oez5	oez6	oiz1	oiz2	oiz3
oiz4	oiz5	oiz6	ooz1	ooz2	ooz5	oz1	oz2
oz3	oz4	oz5	oz6	p	ph	q	r
s	t	th	tr	u1az1	u1az2	u1az3	u1az4
u1az5	u1az6	u1iz4	u1o2iz1	u1o2iz2	u1o2iz3	u1o2iz4	u1o2iz5
u1o2iz6	u1o2uz1	u1o2uz5	u1o2uz6	u1o2z1	u1o2z2	u1o2z3	u1o2z4
u1o2z5	u1o2z6	u1uz1	u1uz2	u1uz3	u1uz4	u1uz5	u1uz6
u1z1	u1z2	u1z3	u1z4	u1z5	u1z6	ua1z1	ua1z2
ua1z4	ua1z5	ua1z6	ua2yz1	ua2yz2	ua2yz3	ua2yz4	ua2yz5
ua2yz6	ua2z1	ua2z2	ua2z3	ua2z4	ua2z5	ua2z6	uaiz1
uaiz2	uaiz5	uaiz6	uaoz5	uaoz6	uauz6	uayz1	uayz2
uayz4	uayz6	uaz1	uaz2	uaz3	uaz4	uaz5	uaz6
ue1z1	ue1z2	ue1z4	ue1z5	ue1z6	ueoz1	ueoz5	ueoz6
uez1	uez2	uez3	uez4	uez5	uez6	uiz1	uiz2
uiz3	uiz4	uiz5	uiz6	uo1iz1	uo1iz2	uo1iz3	uo1iz4
uo1iz5	uo1iz6	uo1z1	uo1z2	uo1z3	uo1z4	uo1z5	uo1z6
uo2iz5	uo2z1	uo2z2	uo2z3	uo2z4	uo2z5	uyaz1	uye1z1
uye1z2	uye1z3	uye1z4	uye1z5	uye1z6	uyuz4	uyuz6	uyz1
uyz2	uyz3	uyz4	uyz5	uyz6	uz1	uz2	uz3
uz4	uz5	uz6	v	x	ye1uz1	ye1uz4	ye1uz5
ye1z1	ye1z4	ye1z5	yz1	yz2	yz3	yz4	yz5
yz6							

A.3. Phoneme set of rhyme-based strategy

Tab. A.19: Phone-based phoneme set: C1R

aa	aach	aaih	aak	aam	aan
aang	aanh	aap	aat	aaw	ahih
ahk	ahm	ahn	ahng	ahp	aht
ahw	ao	aoih	aok	aom	aon
aong	aop	aot	auih	auk	aum
aun	aung	aup	aut	auw	ax
axih	axm	axn	axp	axt	b
d	ee	eech	eem	een	eenh
eep	eet	eew	eh	ehk	ehm
ehn	ehng	ehp	eht	ehw	f
g	h	ch	ie	iek	iem
ien	ieng	iep	iet	iew	iy
iych	iym	iyn	iynh	iyp	iyt
iyw	k	kh	l	m	n
ng	nh	oo	ooih	ook	oom
oon	oong	oop	oot	r	s
sh	t	th	tr	ua	uaih
uak	uam	uan	uang	uap	uat
uaw	uh	uhih	uhk	uhm	uhng
uht	uhw	uo	uoih	uok	uom
uon	uong	uot	uw	uwk	uwm
uwn	uwng	uwp	uwt	v	waa
waach	waaih	waak	waam	waan	waang
waanh	waat	waaw	wahih	wahn	wahng
waht	waok	wauih	wauk	waum	waun
waung	waup	waut	wauw	wax	waxih
waxn	waxt	wee	weech	ween	weet
weh	wehn	wehp	weht	wehw	wie
wien	wiet	wiy	wiych	wiyn	wiynh
wiyt	wiyw	y	zh		

Tab. A.20: Grapheme-based phoneme set: C1R

a	a1c	a1m	a1n	a1ng	a1p
a1t	a2c	a2m	a2n	a2ng	a2p
a2t	a2u	a2y	ac	ach	ai
am	an	ang	anh	ao	ap
at	au	ay	b	c	d
d1	e	e1	e1ch	e1m	e1n
e1nh	e1p	e1t	e1u	ec	em
en	eng	eo	ep	et	g
gh	gi	h	ch	i	ia
ie1c	ie1m	ie1n	ie1ng	ie1p	ie1t
ie1u	ich	im	in	inh	ip
it	iu	k	kh	l	m
n	ng	ngh	nh	o	o1
o1c	o1i	o1m	o1n	o1ng	o1p
o1t	o2	o2i	o2m	o2n	o2p
o2t	oa	oa1c	oa1m	oa1n	oa1ng
oa1t	oac	oach	oai	oam	oan
oang	oanh	oao	oat	oay	oc
oe	oen	oeo	oet	oi	om
on	ong	ooc	oong	op	ot
ph	q	r	s	t	th
tr	u	u1	u1a	u1c	u1i
u1m	u1ng	u1o2c	u1o2i	u1o2m	u1o2n
u1o2ng	u1o2p	u1o2t	u1o2u	u1t	u1u
ua	ua1c	ua1m	ua1n	ua1ng	ua1p
ua1t	ua2n	ua2ng	ua2t	ua2y	uac
uach	uai	uan	uang	uanh	uao
uat	uau	uay	uc	ue	ue1
ue1ch	ue1n	ue1t	uen	ueo	uep
uet	ui	uit	um	un	ung
uo1c	uo1i	uo1m	uo1n	uo1ng	uo1t
uo2	uo2i	uo2n	uo2t	up	ut
uy	uya	uye1n	uye1t	uych	uyn
uynh	uyt	uyu	v	x	y
ye1m	ye1n	ye1t	ye1u		

Tab. A.21: Phone-based phoneme set: C1RT_I

aa	aach	aaih	aak	aam	aan
aang	aanh	aap	aat	aaw	ahih
ahk	ahm	ahn	ahng	ahp	aht
ahw	ao	aoih	aok	aom	aon
aong	aop	aot	auih	auk	aum
aun	aung	aup	aut	auw	ax
axih	axm	axn	axp	axt	b
d	ee	eech	eem	een	eenh
eep	eet	eew	eh	ehk	ehm
ehn	ehng	ehp	eht	ehw	f
g	h	ch	ie	iek	iem
ien	ieng	iep	iet	iew	iy
iych	iym	iyn	iynh	iyp	iyt
iyw	k	kh	l	m	n
ng	nh	oo	ooih	ook	oom
oon	oong	oop	oot	r	s
sh	t	th	tr	ua	uaih
uak	uam	uan	uang	uap	uat
uaw	uh	uhih	uhk	uhm	uhng
uht	uhw	uo	uoih	uok	uom
uon	uong	uot	uw	uwk	uwm
uwn	uwng	uwp	uwt	v	waa
waach	waaih	waak	waam	waan	waang
waanh	waat	waaw	wahih	wahn	wahng
waht	waok	wauih	wauk	waum	waun
waung	waup	waut	wauw	wax	waxih
waxn	waxt	wee	weech	ween	weet
weh	wehn	wehp	weht	wehw	wie
wien	wiet	wiy	wiych	wiyn	wiynh
wiyt	wiyw	y	z1	z2	z3
z4	z5	z6	zh		

Tab. A.22: Grapheme-based phoneme set: C1RT_I

a	a1c	a1m	a1n	a1ng	a1p
a1t	a2c	a2m	a2n	a2ng	a2p
a2t	a2u	a2y	ac	ach	ai
am	an	ang	anh	ao	ap
at	au	ay	b	c	d
d1	e	e1	e1ch	e1m	e1n
e1nh	e1p	e1t	e1u	ec	em
en	eng	eo	ep	et	g
gh	gi	h	ch	i	ia
ie1c	ie1m	ie1n	ie1ng	ie1p	ie1t
ie1u	ich	im	in	inh	ip
it	iu	k	kh	l	m
n	ng	ngh	nh	o	o1
o1c	o1i	o1m	o1n	o1ng	o1p
o1t	o2	o2i	o2m	o2n	o2p
o2t	oa	oa1c	oa1m	oa1n	oa1ng
oa1t	oac	oach	oai	oam	oan
oang	oanh	oao	oat	oay	oc
oe	oen	oeo	oet	oi	om
on	ong	ooc	oong	op	ot
ph	q	r	s	t	th
tr	u	u1	u1a	u1c	u1i
u1m	u1ng	u1o2c	u1o2i	u1o2m	u1o2n
u1o2ng	u1o2p	u1o2t	u1o2u	u1t	u1u
ua	ua1c	ua1m	ua1n	ua1ng	ua1p
ua1t	ua2n	ua2ng	ua2t	ua2y	uac
uach	uai	uan	uang	uanh	uao
uat	uau	uay	uc	ue	ue1
ue1ch	ue1n	ue1t	uen	ueo	uep
uet	ui	uit	um	un	ung
uo1c	uo1i	uo1m	uo1n	uo1ng	uo1t
uo2	uo2i	uo2n	uo2t	up	ut
uy	uya	uye1n	uye1t	uych	uyn
uynh	uyt	uyu	v	x	y
ye1m	ye1n	ye1t	ye1u	z1	z2
z3	z4	z5	z6		

Tab. A.23: Phone-based phoneme set: C1RT_D

aachz5	aachz6	aaihz1	aaihz2	aaihz3	aaihz4	aaihz5	aaihz6
aakz5	aakz6	aamz1	aamz2	aamz3	aamz4	aamz5	aamz6
aangz1	aangz2	aangz3	aangz4	aangz5	aangz6	aanhz1	aanhz2
aanhz3	aanhz4	aanhz5	aanhz6	aanz1	aanz2	aanz3	aanz4
aanz5	aanz6	aapz5	aapz6	aatz5	aatz6	aawz1	aawz2
aawz3	aawz4	aawz5	aawz6	aaz1	aaz2	aaz3	aaz4
aaz5	aaz6	ahihz1	ahihz2	ahihz3	ahihz4	ahihz5	ahihz6
ahkz5	ahkz6	ahmz1	ahmz2	ahmz3	ahmz4	ahmz5	ahmz6
ahngz1	ahngz2	ahngz3	ahngz4	ahngz5	ahnz1	ahnz2	ahnz3
ahnz4	ahnz5	ahnz6	ahpz5	ahpz6	ahtz5	ahtz6	ahwz1
ahwz2	ahwz3	ahwz4	ahwz5	ahwz6	aoihz1	aoihz2	aoihz3
aoihz4	aoihz5	aoihz6	aokz5	aokz6	aomz1	aomz2	aomz3
aomz4	aomz5	aomz6	aongz1	aongz2	aongz3	aongz4	aongz5
aongz6	aonz1	aonz2	aonz3	aonz4	aonz5	aonz6	aopz5
aopz6	aotz5	aotz6	aoz1	aoz2	aoz3	aoz4	aoz5
aoz6	auihz1	auihz2	auihz3	auihz4	auihz5	auihz6	aukz5
aukz6	aumz1	aumz2	aumz3	aumz4	aumz5	aumz6	aungz1
aungz2	aungz3	aungz4	aungz5	aungz6	aunz1	aunz2	aunz3
aunz4	aunz5	aunz6	aupz5	aupz6	autz5	autz6	auwz1
auwz2	auwz4	auwz5	auwz6	axihz1	axihz2	axihz3	axihz4
axihz5	axihz6	axmz1	axmz2	axmz3	axmz4	axmz5	axmz6
axnz1	axnz2	axnz3	axnz4	axnz5	axnz6	axpz5	axpz6
axtz5	axtz6	axz1	axz2	axz3	axz4	axz5	axz6
b	d	eechz5	eechz6	eemz1	eemz2	eemz3	eemz4
eemz5	eemz6	eenhz1	eenhz2	eenhz3	eenhz4	eenhz5	eenhz6
eenz1	eenz2	eenz4	eenz5	eenz6	eepz5	eepz6	eetz5
eetz6	eewz1	eewz2	eewz3	eewz4	eewz5	eewz6	eez1
eez2	eez3	eez4	eez5	eez6	ehkz5	ehkz6	ehmz1
ehmz2	ehmz3	ehmz4	ehmz5	ehmz6	ehngz1	ehngz2	ehngz4
ehngz5	ehnz1	ehnz2	ehnz3	ehnz4	ehnz5	ehnz6	ehpz5
ehpz6	ehtz5	ehtz6	ehwz1	ehwz2	ehwz3	ehwz4	ehwz5
ehwz6	ehz1	ehz2	ehz3	ehz4	ehz5	ehz6	f
g	h	ch	iekz5	iekz6	iemz1	iemz2	iemz3
iemz4	iemz5	iemz6	iengz1	iengz2	iengz3	iengz4	iengz5
iengz6	ienz1	ienz2	ienz3	ienz4	ienz5	ienz6	iepz5
iepz6	ietz5	ietz6	iewz1	iewz2	iewz3	iewz4	iewz5
iewz6	iez1	iez2	iez3	iez4	iez5	iez6	iychz5
iychz6	iymz1	iymz2	iymz3	iymz4	iymz5	iymz6	iynhz1
iynhz2	iynhz3	iynhz4	iynhz5	iynhz6	iynz1	iynz2	iynz4
iynz5	iynz6	iypz5	iypz6	iytz5	iytz6	iywz1	iywz2
iywz3	iywz4	iywz5	iywz6	iyz1	iyz2	iyz3	iyz4
iyz5	iyz6	k	kh	l	m	n	ng
nh	ooihz1	ooihz2	ooihz3	ooihz4	ooihz5	ooihz6	ookz5
ookz6	oomz1	oomz2	oomz3	oomz4	oomz5	oongz1	oongz2
oongz3	oongz4	oongz5	oongz6	oonz1	oonz2	oonz3	oonz4
oonz5	oonz6	oopz5	oopz6	ootz5	ootz6	ooz1	ooz2
ooz3	ooz4	ooz5	ooz6	r	s	sh	t

th	tr	uaihz1	uaihz2	uaihz3	uaihz4	uaihz5	uaihz6
uakz5	uakz6	uamz1	uamz2	uamz5	uamz6	uangz1	uangz2
uangz3	uangz4	uangz5	uangz6	uanz1	uanz2	uanz3	uanz5
uanz6	uapz5	uatz5	uatz6	uawz1	uawz5	uawz6	uaz1
uaz2	uaz3	uaz4	uaz5	uaz6	uhihz4	uhkz5	uhkz6
uhmz2	uhngz1	uhngz2	uhngz3	uhngz4	uhngz5	uhngz6	uhtz5
uhtz6	uhwz1	uhwz2	uhwz3	uhwz4	uhwz5	uhwz6	uhz1
uhz2	uhz3	uhz4	uhz5	uhz6	uoihz1	uoihz2	uoihz3
uoihz4	uoihz5	uoihz6	uokz5	uokz6	uomz1	uomz2	uomz3
uomz5	uomz6	uongz1	uongz2	uongz3	uongz4	uongz5	uongz6
uonz1	uonz2	uonz5	uonz6	uotz5	uotz6	uoz1	uoz2
uoz3	uoz4	uoz5	uoz6	uwkz5	uwkz6	uwmz1	uwmz2
uwmz3	uwmz4	uwmz5	uwmz6	uwngz1	uwngz2	uwngz3	uwngz4
uwngz5	uwngz6	uwnz1	uwnz2	uwnz3	uwnz4	uwnz5	uwnz6
uwpz5	uwpz6	uwtz5	uwtz6	uwz1	uwz2	uwz3	uwz4
uwz5	uwz6	v	waachz5	waachz6	waaihz1	waaihz2	waaihz3
waaihz4	waaihz5	waaihz6	waakz5	waakz6	waamz2	waamz6	waangz1
waangz2	waangz3	waangz4	waangz5	waangz6	waanhz1	waanhz2	waanhz4
waanhz5	waanhz6	waanz1	waanz2	waanz3	waanz4	waanz5	waanz6
waatz5	waatz6	waawz5	waawz6	waaz1	waaz2	waaz3	waaz4
waaz5	waaz6	wahihz1	wahihz2	wahihz3	wahihz4	wahihz5	wahihz6
wahngz1	wahngz2	wahnz1	wahnz2	wahnz3	wahnz4	wahnz5	wahnz6
wahtz5	wahtz6	waokz5	wauihz1	wauihz2	wauihz4	wauihz5	wauihz6
waukz5	waukz6	waumz1	waumz2	waumz5	waumz6	waungz1	waungz2
waungz3	waungz4	waungz5	waungz6	waunz1	waunz2	waunz4	waunz5
waunz6	waupz5	wautz5	wautz6	wauwz6	waxihz5	waxnz3	waxtz5
waxz1	waxz2	waxz4	weechz5	weechz6	weenz1	weenz5	weetz6
weez1	weez2	weez4	weez5	weez6	wehnz1	wehnz2	wehnz4
wehpz5	wehtz5	wehtz6	wehwz1	wehwz2	wehwz4	wehwz5	wehwz6
wehz1	wehz2	wehz3	wehz4	wehz5	wehz6	wienz1	wienz2
wienz3	wienz4	wienz5	wienz6	wietz5	wietz6	wiez1	wiychz5
wiychz6	wiynhz1	wiynhz2	wiynhz5	wiynz1	wiytz5	wiytz6	wiywz4
wiywz6	wiyz1	wiyz2	wiyz3	wiyz4	wiyz5	wiyz6	y
zh							

Tab. A.24: Grapheme-based phoneme set: C1RT_D

a1cz5	a1cz6	a1mz1	a1mz2	a1mz3	a1mz4	a1mz5	a1mz6
a1ngz1	a1ngz2	a1ngz3	a1ngz4	a1ngz5	a1ngz6	a1nz1	a1nz2
a1nz3	a1nz4	a1nz5	a1nz6	a1pz5	a1pz6	a1tz5	a1tz6
a2cz5	a2cz6	a2mz1	a2mz2	a2mz3	a2mz4	a2mz5	a2mz6
a2ngz1	a2ngz2	a2ngz3	a2ngz4	a2ngz5	a2nz1	a2nz2	a2nz3
a2nz4	a2nz5	a2nz6	a2pz5	a2pz6	a2tz5	a2tz6	a2uz1
a2uz2	a2uz3	a2uz4	a2uz5	a2uz6	a2yz1	a2yz2	a2yz3
a2yz4	a2yz5	a2yz6	acz5	acz6	achz5	achz6	aiz1
aiz2	aiz3	aiz4	aiz5	aiz6	amz1	amz2	amz3
amz4	amz5	amz6	angz1	angz2	angz3	angz4	angz5
angz6	anhz1	anhz2	anhz3	anhz4	anhz5	anhz6	anz1
anz2	anz3	anz4	anz5	anz6	aoz1	aoz2	aoz3
aoz4	aoz5	aoz6	apz5	apz6	atz5	atz6	auz1
auz2	auz4	auz5	auz6	ayz1	ayz2	ayz3	ayz4
ayz5	ayz6	az1	az2	az3	az4	az5	az6
b	c	d	d1	e1chz5	e1chz6	e1mz1	e1mz2
e1mz3	e1mz4	e1mz5	e1mz6	e1nhz1	e1nhz2	e1nhz3	e1nhz4
e1nhz5	e1nhz6	e1nz1	e1nz2	e1nz4	e1nz5	e1nz6	e1pz5
e1pz6	e1tz5	e1tz6	e1uz1	e1uz2	e1uz3	e1uz4	e1uz5
e1uz6	e1z1	e1z2	e1z3	e1z4	e1z5	e1z6	ecz5
ecz6	emz1	emz2	emz3	emz4	emz5	emz6	engz1
engz2	engz4	engz5	enz1	enz2	enz3	enz4	enz5
enz6	eoz1	eoz2	eoz3	eoz4	eoz5	eoz6	epz5
epz6	etz5	etz6	ez1	ez2	ez3	ez4	ez5
ez6	g	gh	gi	h	ch	iaz1	iaz2
iaz3	iaz4	iaz5	iaz6	ie1cz5	ie1cz6	ie1mz1	ie1mz2
ie1mz3	ie1mz4	ie1mz5	ie1mz6	ie1ngz1	ie1ngz2	ie1ngz3	ie1ngz4
ie1ngz5	ie1ngz6	ie1nz1	ie1nz2	ie1nz3	ie1nz4	ie1nz5	ie1nz6
ie1pz5	ie1pz6	ie1tz5	ie1tz6	ie1uz1	ie1uz2	ie1uz3	ie1uz4
ie1uz5	ie1uz6	ichz5	ichz6	imz1	imz2	imz3	imz4
imz5	imz6	inhz1	inhz2	inhz3	inhz4	inhz5	inhz6
inz1	inz2	inz4	inz5	inz6	ipz5	ipz6	itz5
itz6	iuz1	iuz2	iuz3	iuz4	iuz5	iuz6	iz1
iz2	iz3	iz4	iz5	iz6	k	kh	l
m	n	ng	ngh	nh	o1cz5	o1cz6	o1iz1
o1iz2	o1iz3	o1iz4	o1iz5	o1iz6	o1mz1	o1mz2	o1mz3
o1mz4	o1mz5	o1mz6	o1ngz1	o1ngz2	o1ngz3	o1ngz4	o1ngz5
o1ngz6	o1nz1	o1nz2	o1nz3	o1nz4	o1nz5	o1nz6	o1pz5
o1pz6	o1tz5	o1tz6	o1z1	o1z2	o1z3	o1z4	o1z5
o1z6	o2iz1	o2iz2	o2iz3	o2iz4	o2iz5	o2iz6	o2mz1
o2mz2	o2mz3	o2mz4	o2mz5	o2mz6	o2nz1	o2nz2	o2nz3
o2nz4	o2nz5	o2nz6	o2pz5	o2pz6	o2tz5	o2tz6	o2z1
o2z2	o2z3	o2z4	o2z5	o2z6	oa1cz5	oa1cz6	oa1mz1
oa1mz2	oa1mz5	oa1mz6	oa1ngz2	oa1ngz3	oa1ngz5	oa1nz1	oa1nz2
oa1nz4	oa1nz5	oa1tz5	oa1tz6	oacz5	oacz6	oachz5	oachz6
oaiz1	oaiz2	oaiz3	oaiz4	oaiz5	oaiz6	oamz2	oamz6
oangz1	oangz2	oangz3	oangz4	oangz5	oangz6	oanhz1	oanhz2

oanhz4	oanhz5	oanhz6	oanz1	oanz2	oanz3	oanz4	oanz5
oanz6	oaoz5	oatz5	oatz6	oayz1	oayz4	oayz5	oaz1
oaz2	oaz3	oaz4	oaz5	oaz6	ocz5	ocz6	oenz1
oenz2	oenz4	oeoz1	oeoz2	oeoz4	oeoz5	oeoz6	oetz5
oetz6	oez1	oez2	oez3	oez4	oez5	oez6	oiz1
oiz2	oiz3	oiz4	oiz5	oiz6	omz1	omz2	omz3
omz4	omz5	ongz1	ongz2	ongz3	ongz4	ongz5	ongz6
onz1	onz2	onz3	onz4	onz5	onz6	oocz5	oongz1
oongz2	oongz5	opz5	opz6	otz5	otz6	oz1	oz2
oz3	oz4	oz5	oz6	ph	q	r	s
t	th	tr	u1az1	u1az2	u1az3	u1az4	u1az5
u1az6	u1cz5	u1cz6	u1iz4	u1mz2	u1ngz1	u1ngz2	u1ngz3
u1ngz4	u1ngz5	u1ngz6	u1o2cz5	u1o2cz6	u1o2iz1	u1o2iz2	u1o2iz3
u1o2iz4	u1o2iz5	u1o2iz6	u1o2mz1	u1o2mz2	u1o2mz5	u1o2mz6	u1o2ngz1
u1o2ngz2	u1o2ngz3	u1o2ngz4	u1o2ngz5	u1o2ngz6	u1o2nz1	u1o2nz2	u1o2nz3
u1o2nz5	u1o2nz6	u1o2pz5	u1o2tz5	u1o2tz6	u1o2uz1	u1o2uz5	u1o2uz6
u1tz5	u1tz6	u1uz1	u1uz2	u1uz3	u1uz4	u1uz5	u1uz6
u1z1	u1z2	u1z3	u1z4	u1z5	u1z6	ua1cz5	ua1cz6
ua1mz5	ua1ngz1	ua1ngz4	ua1ngz6	ua1nz1	ua1nz2	ua1nz5	ua1nz6
ua1pz5	ua1tz5	ua1tz6	ua2ngz1	ua2ngz2	ua2nz1	ua2nz2	ua2nz3
ua2nz4	ua2nz5	ua2nz6	ua2tz5	ua2tz6	ua2yz1	ua2yz2	ua2yz3
ua2yz4	ua2yz5	ua2yz6	uacz5	uacz6	uachz5	uachz6	uaiz1
uaiz2	uaiz5	uaiz6	uangz1	uangz2	uangz3	uangz4	uangz5
uangz6	uanhz1	uanhz2	uanhz5	uanhz6	uanz1	uanz2	uanz4
uanz5	uaoz5	uaoz6	uatz5	uatz6	uauz6	uayz1	uayz2
uayz4	uayz6	uaz1	uaz2	uaz3	uaz4	uaz5	uaz6
ucz5	ucz6	ue1chz5	ue1chz6	ue1nz1	ue1nz5	ue1tz6	ue1z1
ue1z2	ue1z4	ue1z5	ue1z6	uenz1	ueoz1	ueoz5	ueoz6
uepz5	uetz5	uetz6	uez1	uez2	uez3	uez4	uitz5
uitz6	uiz1	uiz2	uiz3	uiz4	uiz5	uiz6	umz1
umz2	umz3	umz4	umz5	umz6	ungz1	ungz2	ungz3
ungz4	ungz5	ungz6	unz1	unz2	unz3	unz4	unz5
unz6	uo1cz5	uo1cz6	uo1iz1	uo1iz2	uo1iz3	uo1iz4	uo1iz5
uo1iz6	uo1mz1	uo1mz2	uo1mz3	uo1mz5	uo1mz6	uo1ngz1	uo1ngz2
uo1ngz3	uo1ngz4	uo1ngz5	uo1ngz6	uo1nz1	uo1nz2	uo1nz5	uo1nz6
uo1tz5	uo1tz6	uo2iz5	uo2nz3	uo2tz5	uo2z1	uo2z2	uo2z4
upz5	upz6	utz5	utz6	uyaz1	uye1nz1	uye1nz2	uye1nz3
uye1nz4	uye1nz5	uye1nz6	uye1tz5	uye1tz6	uychz5	uychz6	uynhz1
uynhz2	uynhz5	uynz1	uytz5	uytz6	uyuz4	uyuz6	uyz1
uyz2	uyz3	uyz4	uyz5	uyz6	uz1	uz2	uz3
uz4	uz5	uz6	v	x	ye1mz4	ye1mz5	ye1nz1
ye1nz5	ye1tz5	ye1uz1	ye1uz4	ye1uz5	yz1	yz2	yz3
yz4	yz5	yz6					

Appendix B

Audio visual experiments

B.1. Weight selection for white noise

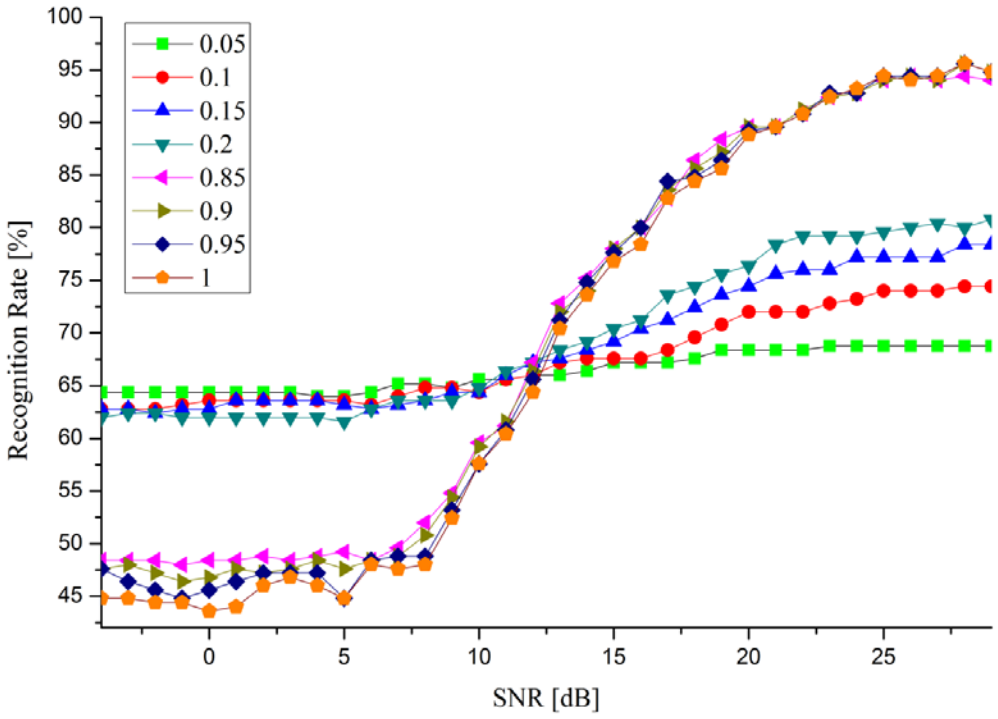


Fig.B.1: LI using N-best dispersion score with different weights *w* in white noise.

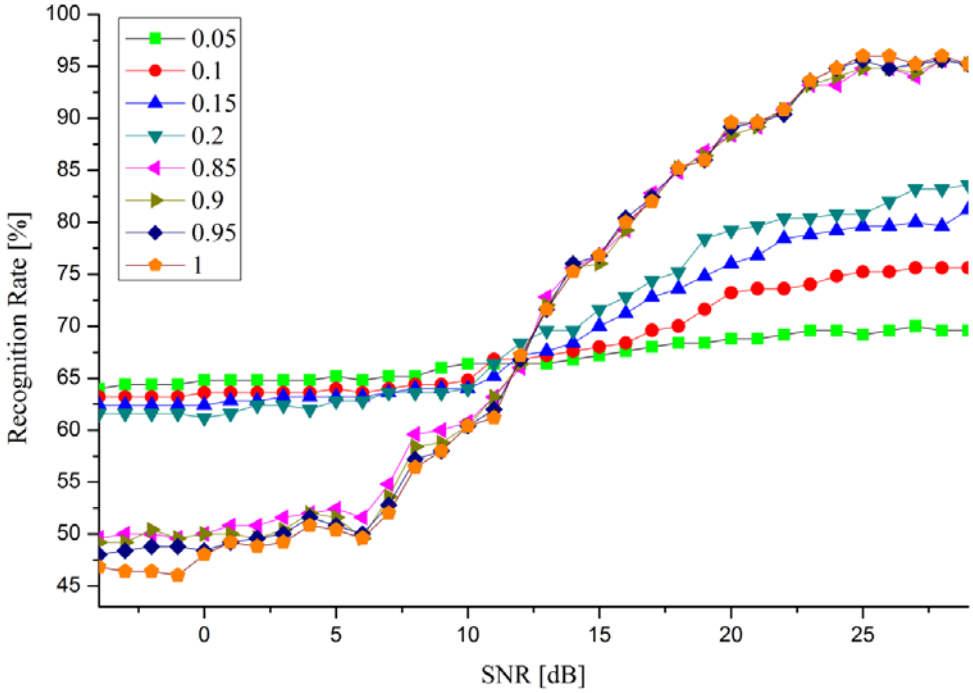


Fig.B.2: LI using Variance score with different weights *w* in white noise.

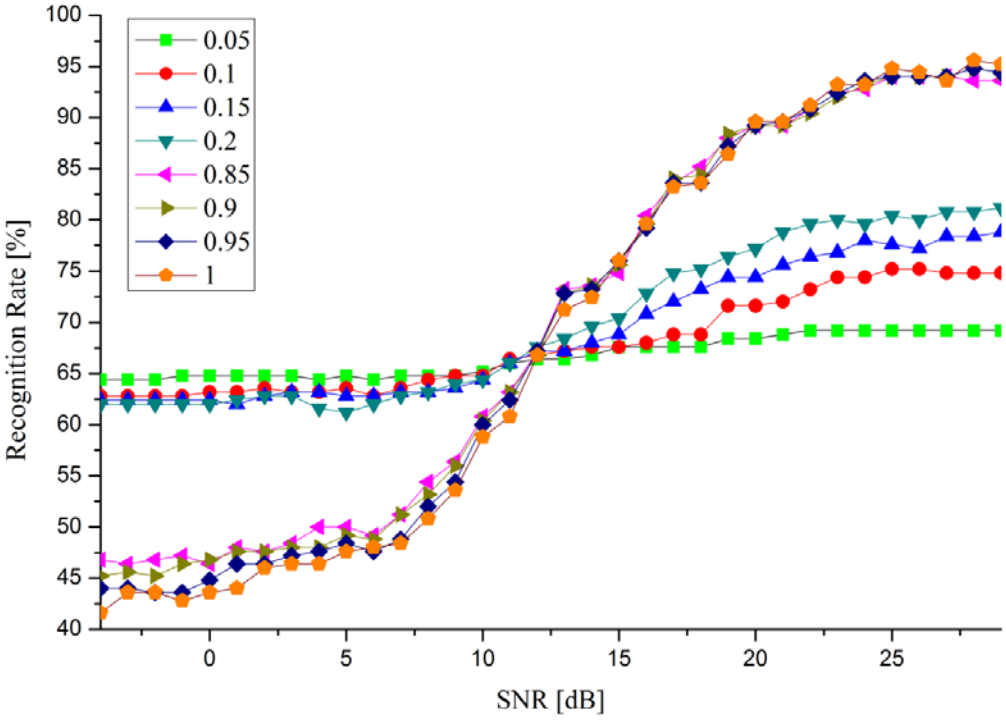


Fig.B.3: LI using N-best average score with different weights *w* in white noise.

B.2. Weight selection for babble noise

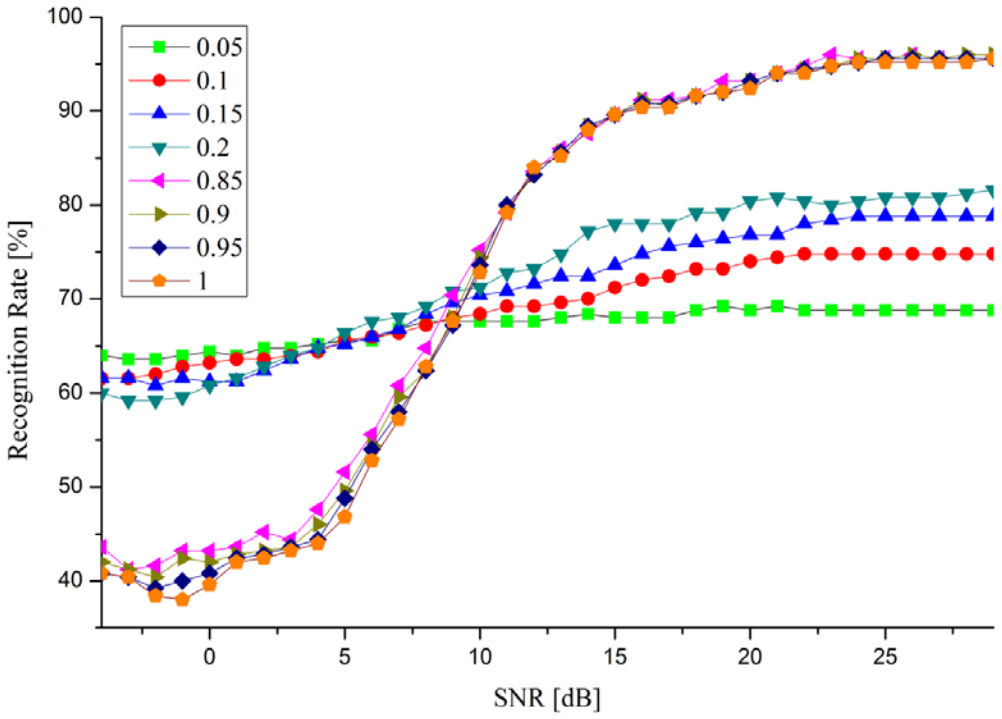


Fig.B.4: LI using N-best dispersion score with different weights *w* in babble noise.

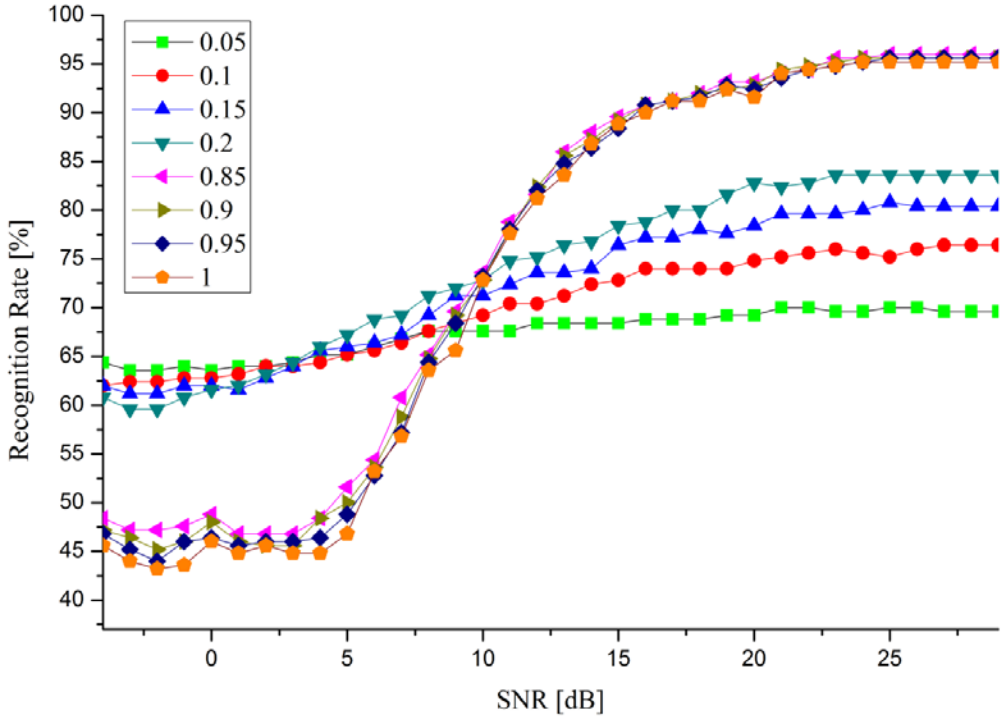


Fig.B.5: LI using Variance score with different weights *w* in babble noise.

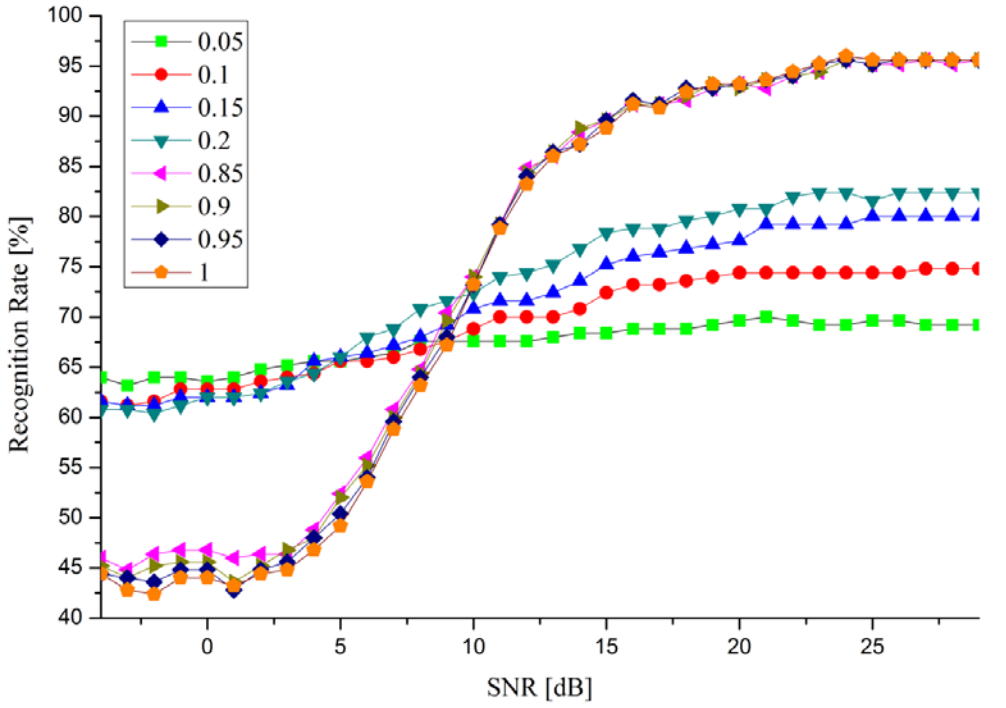


Fig.B.6: LI using N-best average score with different weights *w* in babble noise.

B.3. Weight selection for Volvo noise

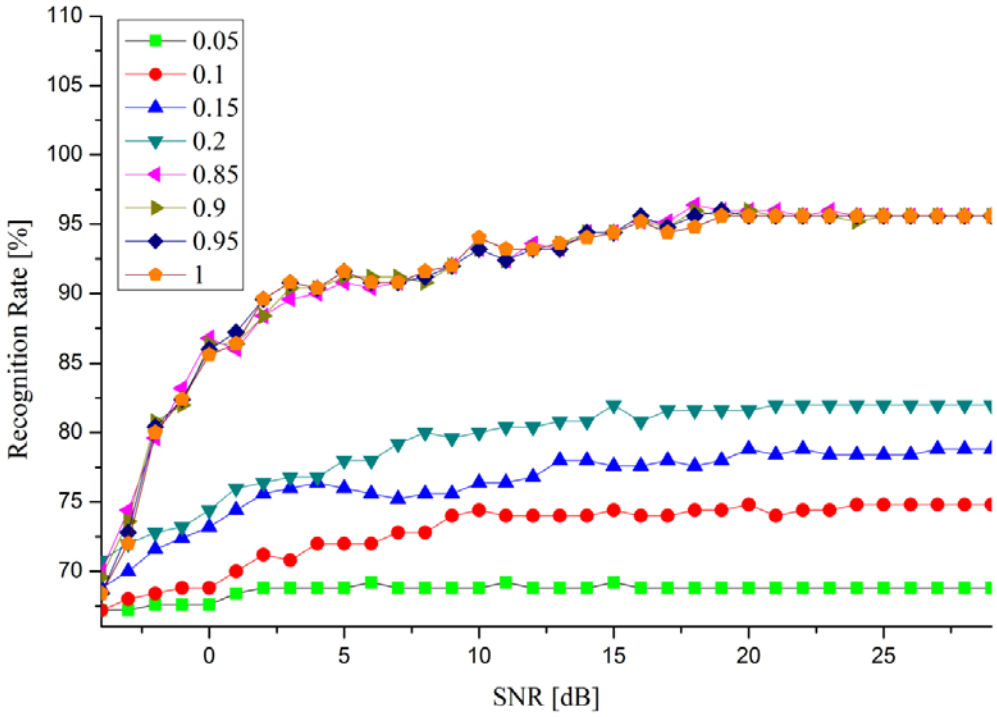


Fig.B.7: LI using N-best dispersion score with different weights *w* in volvo noise.

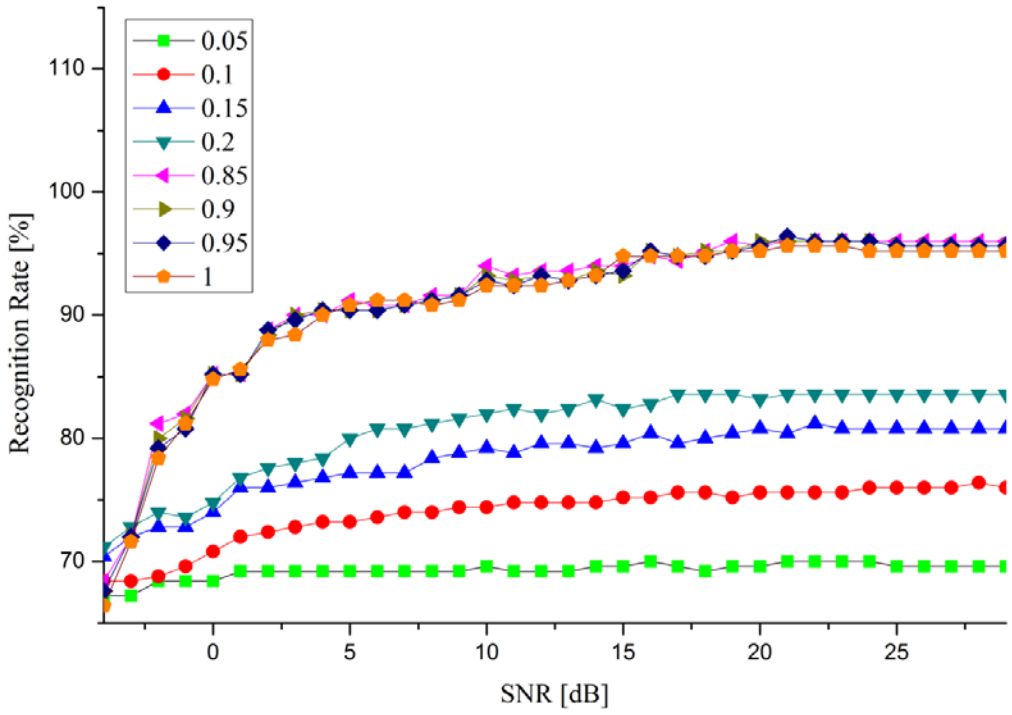


Fig.B.8: LI using Variance score with different weights *w* in volvo noise.

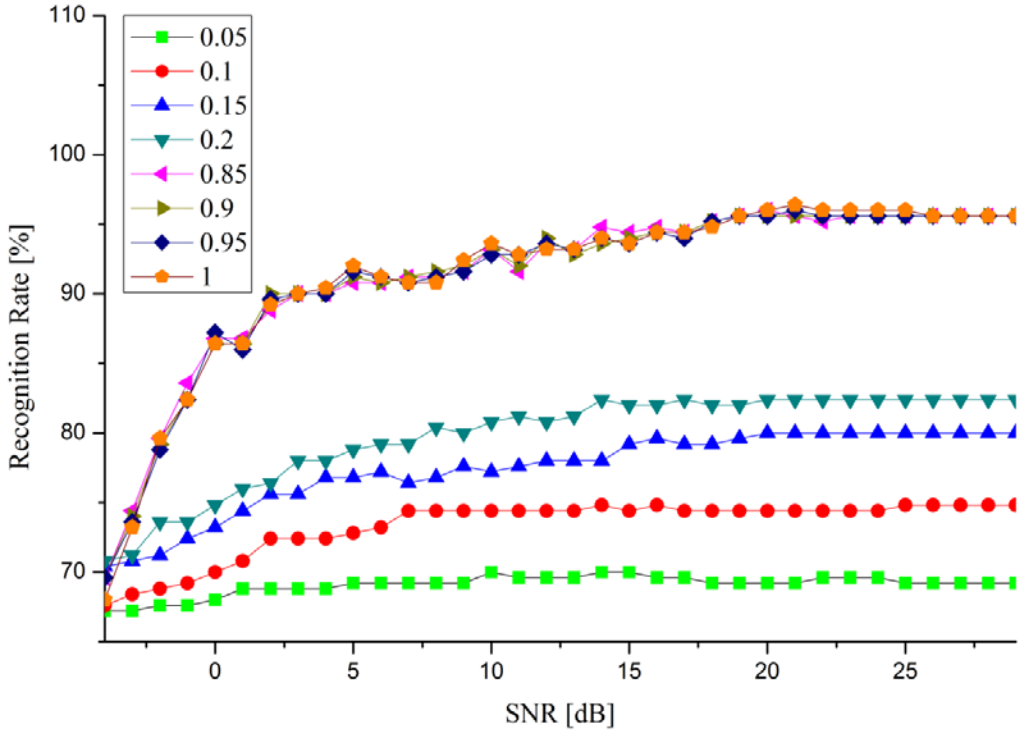


Fig.B.9: LI using N-best average score with different weights *w* in volvo noise.

Bibliography

List of publications

P1. N. T. Chuong and J. Chaloupka, *Phoneme Set and Pronouncing Dictionary Creation for Large Vocabulary Continuous Speech Recognition of Vietnamese*. in *Text, Speech, and Dialogue*, I. Habernal and V. Matoušek, Editors. 2013, Springer Berlin Heidelberg. p. 394-401, Indexed in Scopus Database.

P2. N. T. Chuong and J. Chaloupka, *Developing Text and Speech Databases for Speech Recognition of Vietnamese*. in *IDAACS 2013*, Berlin, Germany, 2013, Indexed in Scopus Database.

P3. N. T. Chuong and J. Chaloupka, *Visual Feature Extraction for Isolated Word Visual Only Speech Recognition of Vietnamese*. in *TSP 2013*, Rome, Italy, 2013, Indexed in Scopus Database.

P4. N. T. Chuong, Selection of sentence set for vietnamese audiovisual corpus design. in *Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2011 IEEE 6th International Conference on*, 2011, pp. 492-495, Indexed in Scopus Database.

P5. N. T. Chuong and J. Chaloupka, *Improvement of Constraint in Active Appearance Model Fitting Algorithm and Its Application in Face Tracking*. in *Proc. of 9th International Workshop on Electronics, Control, Modelling, Measurement and Signals.* Spain, 2009.

List of cited papers

1. Lippmann, R.P., *Speech recognition by machines and humans.* Speech Communication, 1997. 22(1): p. 1-15.

2. Fontaine, V. and H. Bourlard. *Speaker-dependent speech recognition based on phone-like units models-application to voice dialling*. in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. 1997.

3. Yusnita, M.A., et al. *Phoneme-based or isolated-word modeling speech recognition system? An overview*. in *Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on*. 2011.

4. Halavati, R., et al. *A Novel Approach to Very Fast and Noise Robust,Isolated Word Speech Recognition*. in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. 2006.

5. Abushariah, A.A.M., et al. *English digits speech recognition system based on Hidden Markov Models*. in *Computer and Communication Engineering (ICCCE), 2010 International Conference on*. 2010.

6. Aubert, X. and H. Ney. *Large vocabulary continuous speech recognition using word graphs*. in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. 1995.

7. Espy-Wilson and C. Yvonne, *An acoustic-phonetic approach to speech recognition: application to the semivowels*, 1987.

8. Itakura, F., *Minimum prediction residual principle applied to speech recognition.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1975. 23(1): p. 67-72.

9. Rabiner, L., *A tutorial on hidden Markov models and selected applications in speech recognition.* Proceedings of the IEEE, 1989. 77(2): p. 257-286.

10. Myers, C.S. and L.R. Rabiner, *A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition.* Bell System Technical Journal, 1981. 60(7): p. 1389-1409.

11. Nair, N.U. and T.V. Sreenivas. *Multi Pattern Dynamic Time Warping for automatic speech recognition*. in *TENCON 2008 - 2008 IEEE Region 10 Conference*. 2008.

12. Martinez, J., et al. *Speaker recognition using Mel frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques*. in *Electrical Communications and Computers (CONIELECOMP), 2012 22nd International Conference on*. 2012.

13. Bourlard, H., C. Wellekens, and H. Ney. *Connected digit recognition using vector quantization*. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84*. 1984.

14. Bush, M. and G. Kopec. *Network-based connected digit recognition using vector quantization*. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85*. 1985.

15. Makhoul, J., S. Roucos, and H. Gish, *Vector quantization in speech coding.* Proceedings of the IEEE, 1985. 73(11): p. 1551-1588.

16. Rabiner, L. and B.H. Juang, *An introduction to hidden Markov models.* ASSP Magazine, IEEE, 1986. 3(1): p. 4-16.

17. David, E.E. and O.G. Selfridge, *Eyes and Ears for Computers.* Proceedings of the IRE, 1962. 50(5): p. 1093-1101.

18. Davis, K.H., R. Biddulph, and S. Balashek, *Automatic Recognition of Spoken Digits.* The Journal of the Acoustical Society of America, 1952. 24(6): p. 637-642.

19. Olson, H.F. and H. Belar, *Phonetic typewriter.* Audio, IRE Transactions on, 1957. AU-5(4): p. 90-95.

20. Fry, D.B., *Theoretical aspects of mechanical speech recognition.* Radio Engineers, Journal of the British Institution of, 1959. 19(4): p. 211-218.

21. Denes, P., *The design and operation of the mechanical speech recognizer at University College London.* Radio Engineers, Journal of the British Institution of, 1959. 19(4): p. 219-229.

22. Forgie, J.W. and C.D. Forgie, *Results Obtained from a Vowel Recognition Computer Program.* The Journal of the Acoustical Society of America, 1959. 31(11): p. 1480-1489.

23. Suzuki, J. and K. Nakata, *Recognition of Japanese Vowels - Preliminary to the Recognition of Speech.* J.Radio Res.Lab 37(8), 1961: p. 193-212.

24. Sakai, T. and S. Doshita. *The phonetic typewriter, information processing 1962*. in *Proc.IFIP Congress*. 1962.

25. Nagata, K., Y. Kato, and S. Chiba, *Spoken Digit Recognizer for the Japanese Language.* NEC Res.Develop.,No.6, 1963.

26. T.B. Martin, A.L. Nelson, and H.J. Zadell, *Speech Recognition by Feature Abstraction Techniques*, in *Tech.Report AL-TDR-64-176, Air Force Avionics Lab*1964.

27. Vintsyuk, T.K., *Speech discrimination by dynamic programming.* Cybernetics, 1968. 4(1): p. 52-57.

28. Sakoe, H. and S. Chiba, *Dynamic programming algorithm optimization for spoken word recognition.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1978. 26(1): p. 43-49.

29. Myers, C., L. Rabiner, and A.E. Rosenberg, *Performance tradeoffs in dynamic time warping algorithms for isolated word recognition.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1980. 28(6): p. 623-635.

30. Bridle, J., M. Brown, and R. Chamberlain. *An algorithm for connected word recognition*. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82*. 1982.

31. Ney, H., *The use of a one-stage dynamic programming algorithm for connected word recognition.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1984. 32(2): p. 263-271.

32. Reddy, D.R., *Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave.* The Journal of the Acoustical Society of America, 1966. 40(5): p. 1273-1273.

33. Velichko, V.M. and N.G. Zagoruyko, *Automatic recognition of 200 words.* International Journal of Man-Machine Studies, 1970. 2(3): p. 223-234.

34. C.C. Tappert, et al., *Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recover*, 1971: Rome Air Dev.Cen, Rome, NY,Tech.Report TR-71-146.

35. Jelinek, F., L. Bahl, and R. Mercer, *Design of a linguistic statistical decoder for the recognition of continuous speech.* Information Theory, IEEE Transactions on, 1975. 21(3): p. 250-256.

36. Jelinek, F., *The development of an experimental discrete dictation recognizer*. Proceedings of the IEEE, 1985. 73(11): p. 1616-1624.

37. Rabiner, L., et al. *Speaker independent recognition of isolated words using clustering techniques*. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79*. 1979.

38. Klatt, D.H., *Review of the ARPA Speech Understanding Project.* The Journal of the Acoustical Society of America, 1977. 62(6): p. 1345-1366.

39. Lowerre, B., *The Harpy speech understanding system*, in *Readings in speech recognition*, W. Alex and L. Kai-Fu, Editors. 1990, Morgan Kaufmann Publishers Inc. p. 576-586.

40. Sakoe, H., *Two-level DP-matching--A dynamic programming-based pattern matching algorithm for connected word recognition.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1979. 27(6): p. 588-595.

41. Bridle, J.S., M.D. Brown, and R.M. Chamberlain, *Continuous connected word recognition using whole word templates.* Radio and Electronic Engineer, 1983. 53(4): p. 167-175.

42. Myers, C. and L. Rabiner, *A level building dynamic time warping algorithm for connected word recognition.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1981. 29(2): p. 284-297.

43. Chin-Hui, L. and L. Rabiner, *A frame-synchronous network search algorithm for connected word recognition.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1989. 37(11): p. 1649-1658.

44. Baum, L. and T. Petrie, *Statistical Inference for Probabilistic Functions of Finite State Markov Chains.* The Annals of Mathematical Statistics, 1966. 37(6): p. 1554-1563.

45. Baum, L., et al., *A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains.* The Annals of Mathematical Statistics, 1970. 41(1): p. 164-171.

46. Baker, J., *The DRAGON system--An overview.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1975. 23(1): p. 24-29.

47. Jelinek, F., *Continuous speech recognition by statistical methods.* Proceedings of the IEEE, 1976. 64(4): p. 532-556.

48. Lowerre, B.T., *The harpy speech recognition system*, 1976, Carnegie Mellon University. p. 139.

49. Juang, B.H., *On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition—A Unified View.* AT&T Bell Laboratories Technical Journal, 1984. 63(7): p. 1213-1243.

50. Juang, B.H., *Maximum-Likelihood Estimation for Mixture Multivariate Stochastic Observations of Markov Chains.* AT&T Technical Journal, 1985. 64(6): p. 1235-1249.

51. Levinson, S.E., L.R. Rabiner, and M.M. Sondhi, *An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition.* Bell System Technical Journal, 1983. 62(4): p. 1035-1074.

52. Lippmann, R.P., *An introduction to computing with neural nets.* ASSP Magazine, IEEE, 1987. 4(2): p. 4-22.

53. Makino, S., T. Kawabata, and K. Kido. *Recognition of consonant based on the perceptron model*. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83*. 1983.

54. Waibel, A., et al., *Phoneme recognition using time-delay neural networks.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1989. 37(3): p. 328-339.

55. Lippmann, R.P., *Review of neural networks for speech recognition.* Neural Comput., 1989. 1(1): p. 1-38.

56. Price, P., et al. *The DARPA 1000-word resource management database for continuous speech recognition*. in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. 1988.

57. Lee, K.-F., H.-W. Hon, and R. Reddy, *An overview of the SPHINX speech recognition system.* IEEE Transactions on Acoustics, Speech and Signal Processing, 1990. 38(35-45).

58. Chow, Y., et al. *BYBLOS: The BBN continuous speech recognition system*. in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87*. 1987.

59. Weintraub, M., et al. *Linguistic constraints in hidden Markov model based speech recognition*. in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. 1989.

60. Paul, D.B. *The Lincoln robust continuous speech recognizer*. in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. 1989.

61. Zue, V., et al., *The MIT SUMMIT Speech Recognition system: a progress report*, in *Proceedings of the workshop on Speech and Natural Language*1989, Association for Computational Linguistics: Philadelphia, Pennsylvania. p. 179-189.

62. Lee, C.H., et al., *Acoustic modeling for large vocabulary speech recognition.* Computer Speech & Language, 1990. 4(2): p. 127-165.

63. Biing-Hwang, J. and S. Furui, *Automatic recognition and understanding of spoken language - a first step toward natural human-machine communication.* Proceedings of the IEEE, 2000. 88(8): p. 1142-1165.

64. Wu, C., *Minimum Classification Error (MCE) Approach in Pattern Recognition*, in *Pattern Recognition in Speech and Language Processing*2003, CRC Press.

65. Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech.* The Journal of the Acoustical Society of America, 1990. 87(4): p. 1738-1752.

66. Hermansky, H. and N. Morgan, *RASTA processing of speech.* Speech and Audio Processing, IEEE Transactions on, 1994. 2(4): p. 578-589.

67. Nelson Morgan and H. Hermansky, *RASTA Extensions: Robustness to Additive and Convolutional Noise*, in *Workshop on Speech Processing in Adverse Environments, Cannes, France*1992.

68. Eide, E. and H. Gish. *A parametric approach to vocal tract length normalization*. in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. 1996.

69. Wegmann, S., et al. *Speaker normalization on conversational telephone speech*. in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*. 1996.

70. Lee, C.-H. and J.-L. Gauvain, *Bayesian Adaptive Learning and Map Estimation of HMM*, in *Automatic Speech and Speaker Recognition*, C.-H. Lee, F. Soong, and K. Paliwal, Editors. 1996, Springer US. p. 83-107.

71. Chin-Hui, L., C.H. Lin, and J. Biing-Hwang. *A study on speaker adaptation of continuous density HMM parameters*. in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. 1990.

72. Leggetter, C.J. and P.C. Woodland, *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.* Computer Speech & Language, 1995. 9(2): p. 171-185.

73. Biing-Hwang, J. and S. Katagiri, *Discriminative learning for minimum error classification [pattern recognition]*. Signal Processing, IEEE Transactions on, 1992. 40(12): p. 3043-3054.

74. Rabiner, L.R., et al., *Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities.* AT&T Technical Journal, 1985. 64(6): p. 1211-1234.

75. Young, S.J. and L.L. Chase, *Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes.* Computer Speech & Language, 1998. 12(4): p. 263-279.

76. Woodland, P.C. and S.J. Young. *The HTK Tied-State Continuous Speech Recogniser*. in *Third European Conference on Speech Communication and Technology*. 1993. Berlin, Germany: ISCA.

77. Pallett, D.S., et al., *1993 benchmark tests for the ARPA spoken language program*, in *Proceedings of the workshop on Human Language Technology*1994, Association for Computational Linguistics: Plainsboro, NJ. p. 49-74.

78. Hemphill, C.T., J.J. Godfrey, and G.R. Doddington, *The ATIS spoken language systems pilot corpus*, in *Proceedings of the workshop on Speech and Natural Language*1990, Association for Computational Linguistics: Hidden Valley, Pennsylvania. p. 96-101.

79. Watanabe, S., et al., *Variational bayesian estimation and clustering for speech recognition.* Speech and Audio Processing, IEEE Transactions on, 2004. 12(4): p. 365-381.

80. Riccardi, G. and D. Hakkani-Tur, *Active learning: theory and applications to automatic speech recognition.* Speech and Audio Processing, IEEE Transactions on, 2005. 13(4): p. 504-511.

81. Wessel, F. and H. Ney, *Unsupervised training of acoustic models for large vocabulary continuous speech recognition.* Speech and Audio Processing, IEEE Transactions on, 2005. 13(1): p. 23-31.

82. Nakamura, M., K. Iwano, and S. Furui. *The Effect of Spectral Space Reduction in Spontaneous Speech on Recognition Performances*. in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. 2007.

83. Furui, S., et al., *Cluster-based modeling for ubiquitous speech recognition*, in *INTERSPEECH*2005, ISCA. p. 2865-2868.

84. Morris, J. and E. Fosler-Lussier, *Conditional Random Fields for Integrating Local Discriminative Classifiers.* Audio, Speech, and Language Processing, IEEE Transactions on, 2008. 16(3): p. 617-628.

85. Hermansky, H., D.P.W. Ellis, and S. Sharma. *Tandem connectionist feature extraction for conventional HMM systems*. in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*. 2000.

86. Povey, D., et al. *fMPE: Discriminatively Trained Features for Speech Recognition*. in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*. 2005.

87. Yang, L., et al. *Structural metadata research in the EARS program*. in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*. 2005.

88. Soltau, H., et al. *The IBM 2004 conversational telephony system for rich transcription*. in *Proc. ICASSP '05*. 2005.

89. Walker, W., et al., *Sphinx-4: A Flexible Open Source Framework for Speech Recognition.* Sun Microsystems Technical Report, 2004(TR-2004-139).

90. Pham, A.H., *Vietnamese Tone - A New Analysis - Outstanding Dissertations in Linguistics*, 2003: New York: Routledge.

91. Han and Mieko, *Vietnamese vowels*. Studies in the phonology of Asian languages. Vol. 4. 1966, Los Angeles: Acoustic Phonetics Research Laboratory: University of Southern California.

92. Thompson, L.C., *A Vietnamese Reference Grammar* ed. M.-K. Studies1988: University of Hawaii Press.

93. Liêm, N.Đ., *Vietnamese pronunciation*1970: PALI language texts: Southeast Asia., Honolulu: University of Hawaii Press.

94. Hoà, N.Đ., *Vietnamese: Tiếng Việt không son phấn*1997, Amsterdam: John Benjamins Publishing Company.

95. Marc, B. *Coarticulation effects in northern Vietnamese tones*. in *15th International Conference of Phonetic Sciences*. 2003.

96. Lợi, N.V. and E.J. A, *Tones and voice quality in modern northern Vietnamese: Instrumental case studies*, in *Mon-Khmer Studies 28*1998. p. 1–18.

97. Hoang, P., *Syllable Dictionary*1996, Vietnam: Danang publisher.

98. V.B. Le, et al., *Spoken and Written Language Resources for Vietnamese*, in *LREC'04*May 2004: Lisbon, Portugal. p. 599-602.

99. Le, V.B. and L. Besacier, *Comparison of Acoustic Modeling Techniques for Vietnamese and Khmer ASR*, in *ICSLP'06*September 2006: Pittsburgh, PA.

100. Le, V.B., et al., *First steps in building a large vocabulary continuous speech recognition system for Vietnamese,* in *RIVF* February 2005: Can Tho, Vietnam.

101. Le, V.B. and L. Besacier, *First Steps in Fast Acoustic Modeling for A New Target Language: Application to Vietnamese*, in *ICASSP'05*March 2005: Philadelphia, PA.

102. al., V.B.L.e., *"Using the Web for Fast Language Model Construction in Minority Languages"*, in *Eurospeech'03*September 2003: Geneva, Switzerland.

103. H.Q. Nguyen, et al., *A Novel Approach in Continuous Speech Recognition for Vietnamese, an isolating tonal language*, in *Interspeech 2008: 9th Annual Conference of the International Speech Communication Association 2008*2008: Hanoi, Vietnam. p. 1149-1152.

104. Nguyen, H.Q., et al., *Using tone information for Vietnamese continuous speech recognition*, in *RIVF 2008*. p. 103-106.

105. Quang, N.H., et al., *Large vocabulary continuous speech recognition for Vietnamese, an under-resourced language*, in *SLTU 200*2008: Hà Nội, Việt Nam.

106. Le, V.B., et al., *Recent Advances in Automatic Speech Recognition for Vietnamese*, in *Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU 2008)*2008.

107. Le, V.B., et al., *Word sub-word lattices decomposition and combination for speech recognition*, in *International Conference on Acoustics Speech and Signal Processing*2008 IEEE

108. Ha, N., et al. *Progress in Transcription of Vietnamese Broadcast News*. in *First International Conference on Communications and Electronics, 2006. ICCE '06. .* 2006.

109. Tuan, N. and V. Quan. *Advances in Acoustic Modeling for Vietnamese LVCSR*. in *Asian Language Processing, 2009. IALP '09. International Conference on*. 2009.

110. Vu, Q., K. Demuynck, and D. Van Compernolle, *Vietnamese automatic speech recognition: The FLaVoR approach.* Chinese Spoken Language Processing, Proceedings, 2006. 4274: p. 464-474.

111. Pham, N.M., D.A. Duong, and Q.H. Vu. *A robust transcription system for soccer video database*. in *Audio Language and Image Processing (ICALIP), 2010 International Conference on*. 2010.

112. Nguyen, H.Q., V.L. Trinh, and L.T. Dat, *Automatic Speech Recognition for Vietnamese Using HTK System*, in *Computing and Communication Technologies,*

*Research, Innovation, and Vision for the Future (RIVF), 2010 IEEE RIVF International Conference on*2010. p. 1-4.

113. Vu, N.T. and T. Schultz, *Vietnamese Large Vocabulary Continuous Speech Recognition.* 2009 Ieee Workshop on Automatic Speech Recognition & Understanding (Asru 2009), 2009: p. 333-338.

114. Vu, N.T. and T. Schultz, *Optimization On Vietnamese Large Vocabulary Speech Recognition. Workshop on Spoken Languages Technologies for Under-Resourced Languages*, in *SLTU 2010*May 2010: Penang, Malaysia.

115. V.H. Quan, et al., *A generic approach for the Vietnamese handwritten and speech recognition problems.* Developments in Applied Artificail Intelligence, Proceedings, 2002. 2358: p. 47-56.

116. V.H. Quan, P.N. Trung, and D.H.H. Nguyen, *A robust method for the Vietnamese handwritten and speech recognition.* 16th International Conference on Pattern Recognition, Vol Iii, Proceedings, 2002: p. 732-735.

117. T.N. Phung and Q.V. Thai, *A Novel Fast Noise Robust Vietnamese Speech Recognition Applied for Robot Control.* 2008 10th International Conference on Control Automation Robotics & Vision: Icarv 2008, Vols 1-4, 2008: p. 821-826.

118. N.Q. Trung and P.T. Nghia, *The perceptual wavelet feature for noise robust Vietnamese speech recognition.* 2008 Second International Conference on Communications and Electronics, 2008: p. 255-258.

119. Vo Dinh Minh, N. and L. Sungyoung. *Dynamic segmental vector quantization in isolated-word speech recognition*. in *Signal Processing and Information Technology, 2004. Proceedings of the Fourth IEEE International Symposium on*. 2004.

120. Cường, N.Q., P.T.N. Yến, and C. Eric. *Shape vector characterization of Vietnamese tone and application to automatic recognition*. in *Automatic Speech Recognition and Understanding*. 2001. Italia.

121. Duc, D.N., J.-P. Hosom, and C.M. Luong, *"HMM/ANN System for Vietnamese Continuous Digit Recognition"*, in *IEA/AIE*2003. p. 481-486.

122. Nguyễn, H.Q., et al. *Large vocabulary continuous speech recognition for vietnamese, an under-resourced language*. in *SLTU 2008*. 2008. Hà Nội, Việt Nam.

123. Le, V.-B., et al. *Recent advances in automatic speech recognition for vietnamese*. in *Spoken Languages Technologies for Under-Resourced Languages (SLTU 2008)*. 2008.

124. V.B. Le and L. Besacier, *Automatic Speech Recognition for Under-Resourced Languages: Application to Vietnamese Language.* Ieee Transactions on Audio Speech and Language Processing, 2009. 17(8): p. 1471-1482.

125. Ngoc Thang, V. and T. Schultz. *Vietnamese large vocabulary continuous speech recognition*. in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. 2009.

126. Schultz, T. *GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University*. in *International Conference of Spoken Language Processing (ICSLP-2002)*. September 2002. Denver, CO.

127. Thang, T.V., et al. *Vietnamese large vocabulary continuous speech recognition*. in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*. 2005. Lisbon, Portugal: ISCA.

128. Vu, Q., T. Pham, and H. Nguyen, *Towards a Multi-Objective Corpus for Vietnamese Language*, in *COCOSDA2003*2003: Singapore

129. Le, T., H. Nguyen, and Q. VU. *Progress in Transcription of Vietnamese Broadcast News*. in *International Conference on Communications and Electronics (ICCE'06)*. October 2006.

130. Nguyen, T. and Q. Vu, *Advances in Acoustic Modeling for Vietnamese LVCSR*, in *IALP* 2009. p. 280-284.

131. Stolcke, A. *SRILM - an extensible language modeling toolkit*. in *Proceedings of ICSLP*. 2002.

132. Vu, T.T., et al., *Vietnamese large vocabulary continuous speech recognition*, in *INTERSPEECH 2005*2005. p. 1689-1692.

133. Boersma, P. and D. Weenink, *Praat: doing phonetics by computer [Computer program]*. *Version 5.3.59*, 2013.

134. Schubert, K., *Pitch tracking and his application on speech recognition*, 1998, Diploma Thesis at University of Kalsruhe (TH), German.

135. Le, P.N., E. Ambikairajah, and E. Choi. *Improvement of Vietnamese Tone Classification Using FM and MFCC Features*. in *International Conference on Computing and Communication Technologies*. 2009. IEEE-RIVF.

136. N.H. Quang, et al., *"Tone recognition of Vietnamese continuous speech using hidden Markov model"*. 2008 Second International Conference on Communications and Electronics, 2008: p. 233-236.

137. Nocera, P., et al., *Phoneme Lattice Based A* Search Algorithm for Speech Recognition*, in *Text, Speech and Dialogue*, P. Sojka, I. Kopeček, and K. Pala, Editors. 2002, Springer Berlin Heidelberg. p. 301-308.

138. Le, V.B., L. Besacier, and T. Schultz. *Acoustic-Phonetic Unit Similarities For Context Dependent Acoustic Model Portability*. in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. 2006.

139. Young, S., et al., *The HTK Book (for HTK version 3.2)*2002: Cambridge University Engineering Department.

140. Potamianos, G., J. Luettin, and C. Neti. *Hierarchical discriminant features for audio-visual LVCSR*. in *International Conference on Acoustics, Speech and Signal Processing*. 2001. Salt Lake City, UT.

141. Potamianos, G., et al., *A Cascade Visual Front End for Speaker Independent Automatic Speechreading*. International Journal of Speech Technology, 2001. 4(3-4): p. 193-208.

142. Potamianos, G. and C. Neti. *Improved ROI and within frame discriminant features for lipreading*. in *International Conference on Image Processing*. 2001. Thessaloniki, Greece.

143. Rogozan, A., P. Deléglise, and M. Alissali, *Adaptive Determination of Audio and Visual Weights for Automatic Speech Recognition*, 1997: Rhodes.

144. Dupont, S. and J. Luettin, *Audio-Visual Speech Modeling for Continuous Speech Recognition*, in *IEEE TRANSACTIONS ON MULTIMEDIA*SEPTEMBER 2000.

145. Neti, C., et al. *Audio-visual speech recognition*. in *Final Workshop 2000 Report, Center for Language and Speech Processing*. 2000. Baltimore.

146. Matthews, I., et al. *A comparison of model and transform-based visual features for audio-visual LVCSR*. in *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*. 2001.

147. Senior, A.W., *Face and feature finding for a face recognition system*, in *IN SECOND INTERNATIONAL CONFERENCE ON AUDIO- AND VIDEO-BASED BIOMETRIC PERSON AUTHENTICATION*1999. p. 154-159.

148. Chiou, G.I. and J.N. Hwang, *Lipreading from Color Video*, in *IEEE Transactions on Image Processing*1997. p. 1192-1195.

149. Chan, M.T., Y. Zhang, and T.S. Huang. *Real-time lip tracking and bimodal continuous speech recognition*. in *Workshop Multimedia Signal Processing*. 1998.

150. T. Saitoh, e.a. *Analysis of efficient lip reading method for various languages*. in *International Conference on Pattern Recognition (ICPR2008)*. 2008.

151. Komai, Y., Y. Ariki, and T. Takiguchi, *Audio-Visual Speech Recognition Based on AAM Parameter and Phoneme Analysis of Visual Feature*, in *Advances in Image and Video Technology*, Y.-S. Ho, Editor 2012, Springer Berlin Heidelberg. p. 97-108.

152. Potamianos, G., H.P. Graf, and E. Cosatto. *An Image Transform Approach for HMM Based Automatic Lipreading*. in *International Conference on Image Processing*. 1998. Chicago.

153. Shin, J., J. Lee, and D. Kim, *Real-time lip reading system for isolated Korean word recognition.* Pattern Recogn., 2011. 44(3): p. 559-571.

154. Yang, W., S. Lucey, and J.F. Cohn. *Enforcing convexity for improved alignment with constrained local models*. in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008.

155. Lucey, S., et al., *Efficient constrained local model fitting for non-rigid face alignment.* Image Vision Comput., 2009. 27(12): p. 1804-1813.

156. Cristinacce, D. and T. Cootes, *Automatic feature localisation with constrained local models.* Pattern Recognition, 2008. 41(10): p. 3054-3067.

157. Chaloupka, J., J. Nouza, and J. Zdansky. *Audio-visual voice command recognition in noisy conditions*. in *AVSP-2008*. 2008.

158. Varga, A.P., et al., *The noisex-92 study on the effect of additive noise on automatic speech recognition*, 1992: DRA Speech Research Unit.