TECHNICAL UNIVERSITY OF LIBEREC

Faculty of Mechatronics, Informatics and Interdisciplinary Studies

Automatic speech recognition of Vietnamese

Summary of Doctoral Thesis

2014 **Nguyen Thien Chuong**

TECHNICAL UNIVERSITY OF LIBEREC

Faculty of Mechatronics, Informatics and Interdisciplinary Studies

Automatic speech recognition of Vietnamese

Author:	Nguyen Thien Chuong
Supervisor:	doc. Ing. Josef Chaloupka, Ph.D.
Ph.D. Program:	P 2612 Electronics and Informatics
Branch of Study:	2612V045 Technical Cybernetics
Year:	2014

Scope of doctoral thesis

Number of pages: 180

Number of appendixs: 2 + CD

Number of figures: 42

Number of tables: 97

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 of Vietnamese with its corresponding grapheme-to-phoneme mapping table is proposed. By constructing various types of pronouncing dictionaries and language models, the effects of 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 studies are 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 channel. 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

1 Introduction

The researches on automatic speech recognition (ASR) of Vietnamese have made significant progresses since it was first introduced more than twenty years ago. However, ASR of Vietnamese is just at its experiment stage and yet to reach the performance level required to be widely used in real-life applications.

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

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 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 an ASR system. This thesis is mainly concerned with the following basic problems:

The first problem is the proposal of a phoneme set. In this work, both grapheme-based and phone-based phoneme set are proposed and evaluated in the form of LVCSR tasks.

The second problem is the construction of a pronouncing dictionary. 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. All main hypotheses 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, 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.

2 Vietnamese Language

Vietnamese is a Viet-Muong language in the Mon-Khmer group within the Austro-Asiatic family. 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 writing system, Vietnamese language uses a set of Latin symbols. It includes 22 out of 26 letters as in the English alphabet, seven letters which used only in Vietnamese and an addition of six diacritics for tones (Fig. 1). Note that Vietnamese does not have letters “f”, “j”, “w”, and “z” as in English, although they may appear in loanwords.

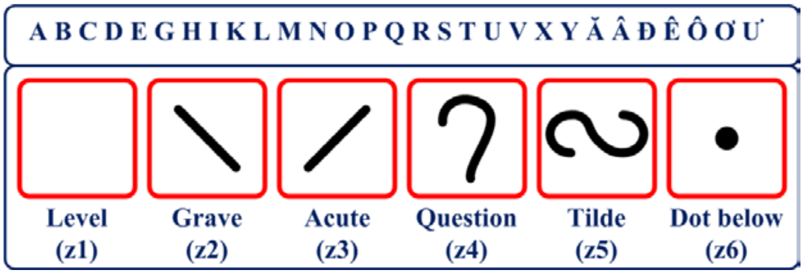

Fig. 1: The Vietnamese alphabet and tone in writing system.

In Vietnamese modern language, there are about 8,000 syllables. Each syllable is modeled as in Fig. 2 and follows the scheme:

$$[\mathbf{C1}]\mathbf{R} + \mathbf{T} \text{ or } [\mathbf{C1}][\mathbf{w}]\mathbf{V}[\mathbf{C2}] + \mathbf{T}$$

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.

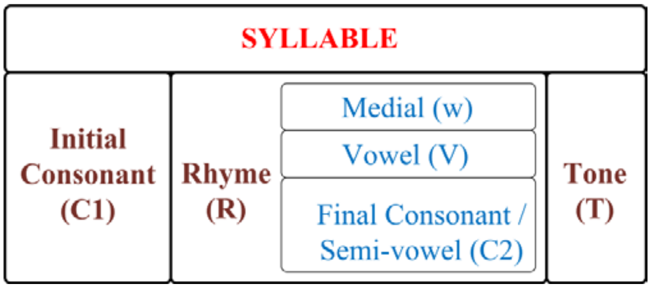

Fig. 2: Vietnamese syllable structure.

3 Strategies for Speech Recognition of Vietnamese

3.1 Phoneme Set Proposal

For ASR of Vietnamese, there are two basic phoneme set types: phone-based phoneme set and grapheme-based phoneme set. 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. 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.

3.2 Pronouncing Dictionary Creation

When constructing a pronouncing 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 pronouncing dictionary including phone-based and grapheme-based pronouncing dictionary. Also, Tab. 1 shows four schemes that are used to analyze a Vietnamese syllable.

Tab. 1: 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. 3 shows various pronouncing dictionary types which are created using different schemes of analysis a Vietnamese syllable and different hypotheses about tone, and Tab. 4 shows the number of basic phonetic units of each dictionary type.

Tab. 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
	t^h	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	

Tab. 3: Vietnamese pronouncing 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. 4: 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 pronouncing dictionary, the procedure 3.1 is used. Note that, for grapheme-based dictionary, step 3 is omitted.

Procedure 3.1: 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.

3.3 Strategies for Speech Recognition of Vietnamese

3.3.1 Phoneme-based strategy

The basic phonetic units to build acoustic models are phonemes: *C1*, *w*, *V* and *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. 4). Note that, the number of basic phonetic units of these methods is also smaller than the corresponding methods 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.

3.3.2 Vowel-based strategy

The basic phonetic units to build acoustic models are *C1*, *M* and *C*. This strategy consists of five basic methods for each of the phone-based and grapheme-based phoneme set in which methods with dependent tone always have larger number of basic phonetic units than the methods with independent tone or without using tone. 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 unit in each syllable is small and the data needed to cover, say all possible bi-phone or tri-phone based ASR system is large.

3.3.3 Rhyme-based strategy

The basic phonetic units to build acoustic models are *C1* and *R*. 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. 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.

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

4 Text, Audio and Audio-Visual Databases

4.1 Building of Vietnamese Text Corpus from the Internet

4.1.1 Vietnamese text corpus from Wikipedia

To extract a Vietnamese text corpus from Wikipedia, a Wiki data dump of Vietnamese is needed. From the Wikipedia service, the dump for page articles which contains all articles with useful text is selected. Then, unnecessary articles are filtered out and plain text is extracted from the dump resulted in a Wiki text corpus (Tab. 5).

Tab. 5 : Statistics Of Text Corpus From Wikipedia

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

4.1.2 Extracting of general purpose text corpus

To build a text corpus from the web, a frequency list of words in the Wiki corpus is first created using a lexicon of Vietnamese with 73,901 common words. Then 5000 seed words are collected from that list to generate web queries of these seed words. Using the following algorithm we obtained the best query length for Vietnamese is 2.

Algorithm 4.1: 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
```

Obtaining the query length, around 35,000 queries are generated using 5,000 seed words. Given these queries, only the URLs which size from 5 KB to 2 MB are downloaded. By extracting all useful text from the web pages, we obtain a text corpus that is large and diverse enough for multipurpose research. Tab. 6 shows some statistics of the corpus.

Tab. 6: Statistics of raw data of general purpose text corpus.

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

4.1.3 Extracting of specific purpose text corpus

In this section, a specific purpose text corpus is created with text mainly in the field of news and literature by collecting several Vietnamese websites which are rich of newspaper and book resource. For each website, only 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.

4.1.4 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. Tab. 7 to Tab. 10 show some statistics of all the filtered text corpora.

Tab. 7: 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. 8: 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. 9: 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. 10: 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

4.2 Collecting of Speech Corpus for LVCSR Task

The speech corpus for LVCSR of Vietnamese is collected from the Internet resource. 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 is 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. This corpus contains speech mainly of type: story reading, news report, weather forecast, and conversation.

4.3 Designing of Audio-Visual Speech Corpus

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 resulted in a total 2500 isolated word, 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.

5 Audio Speech Recognition of Vietnamese

5.1 Building language model for LVCSR of Vietnamese

5.1.1 Syllable-based LM construction

First, to evaluate the effect of different text corpus categories, all the text corpora are utilized. The constructed LMs will have the vocabularies of size 6000 and 7000 syllables. Also, another vocabulary that contains all 5741 distinct syllables occurring in the training part of the speech corpus’s transcription is also 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. The perplexities of all LMs are shown in Tab. 11 to Tab. 14.

To further examine the effect of vocabulary size and smoothing methods on LM constructing, the *VN only* text corpus presented in Tab. 10 is used for LM estimation and the same testing data as the previous experiment is used for evaluation. All LMs are trained using three different smoothing methods including Good-Turning, Kneser-Ney and Witten-Bell. Tab. 15 to Tab. 18 show the perplexity of constructed LMs.

Tab. 11: LM test on general purpose text corpus.

Vocabulary size	**Perplexity VN only**		**Perplexity VN mix**	
	bi-gram	*tri-gram*	*bi-gram*	*tri-gram*
6000	282.6949	223.1744	333.0480	255.2985
7000	286.0785	226.1346	336.9742	258.6981
5741	289.2655	228.9577	340.9702	262.1823

Tab. 12: LM test on specific text corpus (literature).

Vocabulary size	**Perplexity VN only**		**Perplexity VN mix**	
	bi-gram	*tri-gram*	*bi-gram*	*tri-gram*
6000	271.8294	190.4345	268.5521	203.6228
7000	274.5605	192.6637	271.3496	206.0177
5741	277.3793	194.9775	274.1888	208.4674

Tab. 13: LM test on specific text corpus (news).

Vocabulary size	**Perplexity VN only**		**Perplexity VN mix**	
	bi-gram	*tri-gram*	*bi-gram*	*tri-gram*
6000	386.8472	324.8600	468.4447	373.3483
7000	392.2740	329.8564	475.0951	379.2644
5741	396.8644	334.1689	481.4170	384.9309

Tab. 14: 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*
6000	246.8432	160.3296	298.8425	199.4028	254.9655	154.7797
7000	249.4565	162.3357	302.1896	202.0251	257.6787	156.7561
5741	252.1593	164.4235	305.7720	204.8486	260.5325	158.8546

Tab. 15: 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. 16: 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. 17: 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. 18: 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

5.1.2 Multi-syllable-based LM construction

Word-based LM is more difficult to construct than syllable-based LM. This work uses a simple data-driven approach to segment multi-syllabic token for the tasks of multi-syllable-based LM construction. Tab. 19 shows the perplexities of all the bi-gram LMs estimated with Kneser-Ney smoothing using interpolated model.

Tab. 19: Multi-syllable-based LM test.

N	**Perplexity**	**Vocabulary size**	**Number of actual bi-syllable 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

5.2 Isolated word speech recognition

For experiments on isolated word speech recognition, the audio part of the audio-visual database 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. 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). The recognition results are shown in Tab. 20.

Tab. 20: 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**

5.3 Experiments on LVCSR of Vietnamese

For experiments on LVCSR of Vietnamese, the speech corpus collected from the Internet resource is used. Tab. 21 shows some statistics of this speech data. 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.

Tab. 21: Speech corpus for LVCSR tasks.

	Number of speaker		Number of utterance	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

5.3.1 Examining the Effect of Tone in Vietnamese Syllables

To totally examine the effect of position and role of tone in a syllable, context-independent continuous speech recognizers will be trained and tested using the first three strategies described in section 3.3. For each method in these strategies, the phonetic units are trained with 3 states HMM 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. 10) 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. From the recognition results (Tab. 22), some major conclusions can be made by analyzing this information:

- 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.

- 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.

- 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.

Tab. 22: 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

5.3.2 Context-dependent LVCSR of Vietnamese

In this experiment, the performance of context-dependent HMM 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. Tab. 23 shows the syllable accuracies of all methods. In this 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.

Tab. 23: 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

5.3.3 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 as described in the previous experiment.

5.3.3.1 The effect of text corpus category

For this experiment, the bi-gram LMs shown in Tab. 11 to Tab. 14 are used and the SACC of recognizers using the above LMs are shown in Tab. 24. The results show that, LM constructed from the general purpose text corpus provides really good SACC in comparison with LM constructed from text corpus of type *Literature*. Also, it is interested to see that the combined text corpus of the three text categories gives the best SACC.

Tab. 24: 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**

5.3.3.2 The effect of dictionary size and smoothing method

For this experiment, three different smoothing methods including Good-Turning, Kneser-Ney, and Witten-Bell as well as various vocabulary sizes will be examined. All the bi-gram LMs estimated in the previous section are evaluated (Tab. 15 to Tab. 18). Tab. 25 shows that, recognizers based on LMs estimated with the Kneser-Ney smoothing using interpolated model gives the best SACC. 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.

Tab. 25: 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**

5.3.3.3 The effect of multi-syllable-based LM

For this experiment, bi-gram multi-syllable-based LMs described in section 5.1.2 are used. The recognizers with LMs estimated by Kneser-Ney smoothing using interpolated model will be examined. To estimate and evaluate the recognizers using this type of LM, all the training and testing transcription of the speech corpus will be segmented.

Tab. 26 shows the syllable accuracies of all the recognizers trained using different number of bi-syllabic tokens *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.

Tab. 26: SACC [%] for multi-syllable-based LMs.

Number of bi-syllabic token	**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

5.3.4 Gender-dependent LVCSR of Vietnamese

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 sections. 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.

Tab. 27 shows the syllable accuracy of gender-dependent recognizers using different smoothing methods of LM. It is easy to see that the gender-dependent recognizers do improve the recognition results for all smoothing methods of LM in which Kneser-Ney smoothing with interpolated model gives the best result. The best recognition result of 79.66% is an improvement in comparison with 74.11% of gender-independent recognizers.

Tab. 27: SACC [%] for Gender-dependent LVCSR.

		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

6 Audio-Visual Speech Recognition of Vietnamese

6.1 Isolated word visual only speech recognition

In these experiments, 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.

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. Tab.

28 shows 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. 28: 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. The results shows that LDA matrix trained using phoneme as basic class with tone attached to the last component of the syllable give the best result both in accuracy (57%) and dimensional reduction (8).

In the third experiment, the performance of two visual front ends for feature extraction is compared. Tab. 29 shows the best recognition results for each visual front end. It shows that HLDA outperform the 1-Stage LDA and both visual front ends do improve the DCT only visual feature.

Tab. 29: 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

In this experiment, three types of visual feature are compared. Tab. 30 shows the best accuracy is from DCT and PCA (62%), but DCT obtains better dimensional reduction (16).

Tab. 30: 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

6.2 Isolated word audio-visual speech recognition.

For experiments on audio-visual fusion, all adaption and testing data will be synthesized in noise condition. To add noise to audio signal, the segmental signal to noise ratio (SSRN) is first estimated using the audio-visual database. The audio and visual features extracted from 40 speakers are used as training data. Features extracted from other 5 speakers are used as testing data for the recognition task, and features extracted from the last 5 speakers are used as adapting data.

6.2.1.1 Middle integration experiments

For the experiments, three types of noise are used as additive noises including white noise, babble noise and volvo noise. The results shows that in noise condition, the recognition rate for visual only recognizer (VO) is 62.2% and does not change when noise change. But for audio only recognizer (AO), 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.3 to Fig.5 show the results of audio-visual fusion using MI strategy with both weights equal to 1 (MI w11). Fig.3 also shows MI using adaptation data (MI WA) to determine the best weights for each stream in different SSNR values. It is easy to see that audio-visual fusion using MI do improve the recognition results in noise conditions and MI using adaptation data outperform the case where both stream weights equal to 1.

6.2.1.2 Late integration experiments

In this 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. The best audio weight γ obtained from this step for each SSNR value will be applied on the testing part. Fig.3 to Fig.5 show the recognition results for three different noises using this strategy (LI WA). These figures also show the optimal recognition results (LI ES) where the exhausted search is applied directly on the testing data. 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 and can obtain the results as good as the optimal case (LI ES).

Fig.3 to Fig.5 also 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 (LI CS Opt) outperforms the recognizer using only weight *w* = 1 to compute confident scores.

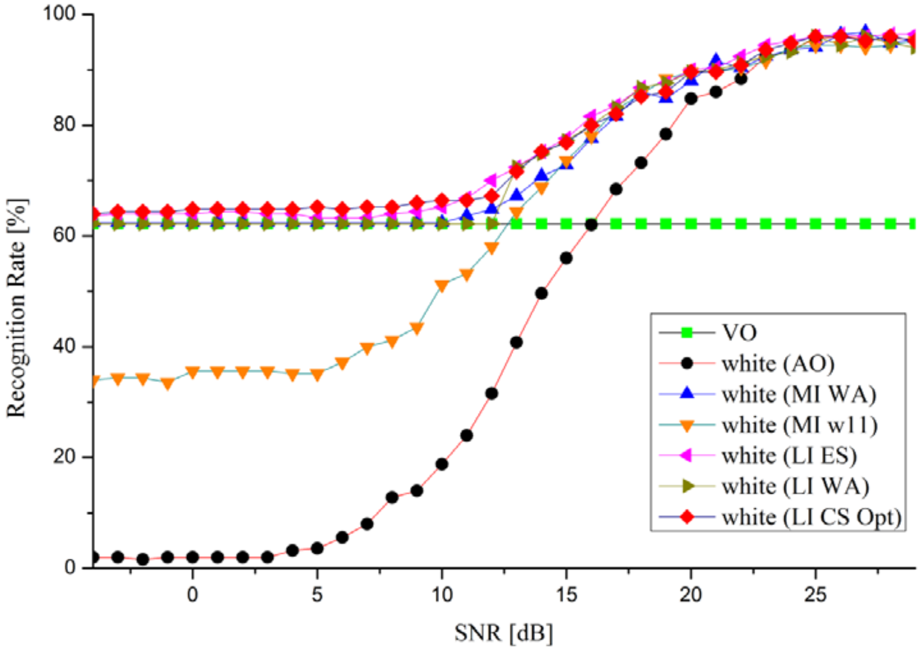

Fig.3: Comparison of fusion strategies in white noise.

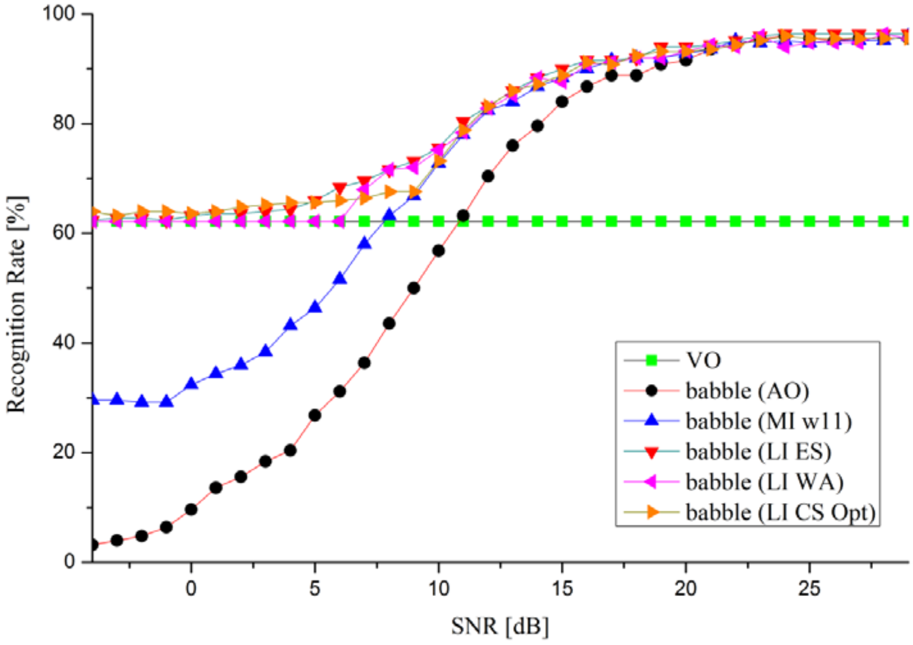

Fig.4: Comparison of fusion strategies in babble noise.

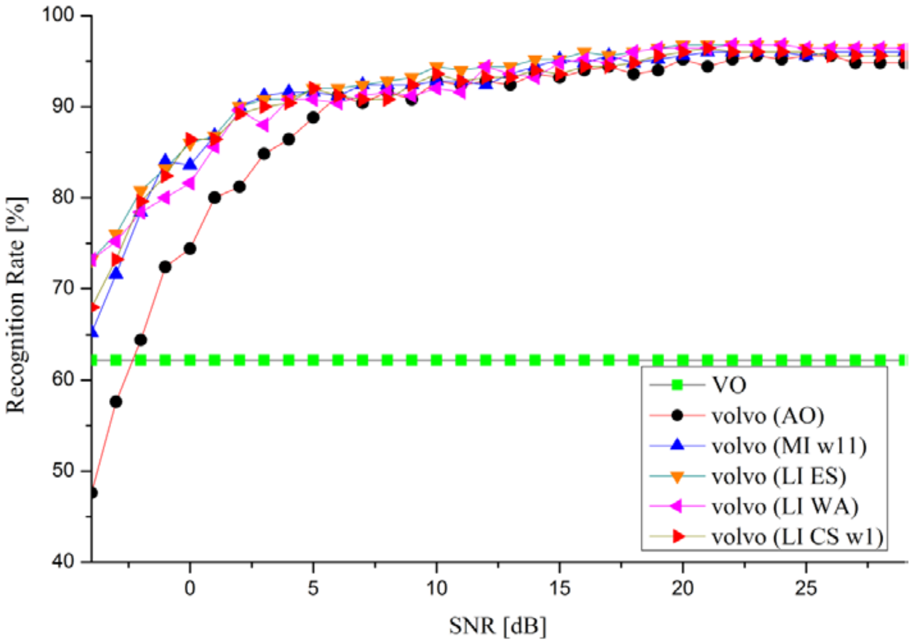

Fig.5: Comparison of fusion strategies in volvo noise.

7 Conclusion

7.1 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. 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.

7.2 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, the recognition accuracies in Tab. 22 showed that 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.

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.

For the final hypothesis, the task of examining the position of tone in Vietnamese syllable is solved. It could be seen that 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.

7.3 LVCSR 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.

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.

7.4 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 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. 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.

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.

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.

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.

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.

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.

Ing. Nguyen Thien Chuong

Automatic speech recognition of Vietnamese
Summary of Doctoral Thesis

Technical University of Liberec
Faculty of Mechatronics, Informatics and Interdisciplinary Studies

30 pages
10 copies

June 2014