Vietnamese Large Vocabulary Continuous Speech Recognition

Ngoc Thang Vu, Tanja Schultz

Cognitive Systems Lab (CSL), Institute for Anthropomatics, University of Karlsruhe

Thang.Vu@student.kit.edu, tanja@ira.uka.de

***Abstract*—We report on our recent efforts toward a large vocabulary Vietnamese speech recognition system. In particular, we describe the Vietnamese text and speech database recently collected as part of our GlobalPhone corpus. The data was complemented by a large collection of text data crawled from various Vietnamese websites. To bootstrap the Vietnamese speech recognition system we used our Rapid Language Adaptation scheme applying a multilingual phone inventory. After initialization we investigated the peculiarities of the Vietnamese language and achieved significant improvements by implementing different tone modeling schemes, extended by pitch extraction, handling multiwords to address the monosyllable structure of Vietnamese, and featuring language modeling based on 5-grams. Furthermore, we addressed the issue of dialectal variations between South and North Vietnam by creating dialect dependent pronunciations and including dialect in the context decision tree of the recognizer. Our currently best recognition system achieves a word error rate of 11.7% on read newspaper speech.**

I. Introduction

In light of the worlds globalization, one of the most important trends in present-day speech technology is the need to support multiple input and output languages, especially when applications are intended for international markets and linguistically diverse user communities. As a result, new algorithms and strategies are required, which support a rapid adaptation of speech processing systems to new languages. Currently, the time and costs associated with this task is one of the major bottlenecks in the development of multilingual speech technology. Our Rapid Language Adaptation Tools (RLAT) [9] aim to significantly reduce the amount of time and effort involved in building speech processing systems for new languages. It is envisioned to be achieved by providing innovative methods and tools that enable users to develop speech processing models, collect appropriate speech and text data to build these models, as well as evaluate the results allowing for iterative improvements. In this paper we describe the application of these tools to the Vietnamese language. Despite recent developments in Vietnam to address speech and language technologies, with groups such as LTI [11] and MICA [7] among others, Vietnamese might still be considered to be one of the under-resourced languages. The language provides many challenges for speech and language processing, such as tonal speech features, its monosyllabic language structure, and dialectal variations. The purpose of this study is to apply our RLAT tools to Vietnamese for collecting a large speech and text corpus, increase our knowledge to Vietnamese speech recognition, and extend language dependent modules in RLAT and the speech recognition system to include the peculiarities of the Vietnamese language. The paper is organized as follows. In Section II we give a brief introduction to the structure of the Vietnamese language. Section III describes our data speech and text data collection which was used to train and evaluate all parts of the Vietnamese speech recognizer. In Section IV we present our baseline recognizer resulting from the rapid initialization based on RLAT. Section V gives a detailed description of the optimization steps performed to address certain challenges of Vietnamese, including the implementation of multiwords, extraction of pitch information, language model improvements, and selective collection of additional speech data for system improvement. Section VI presents the experimental results and includes our findings on handling dialectal variations. We conclude in Section VII with a summary of current results and an outlook to future work.

II. Structure of Vietnamese language

Vietnamese is a monosyllabic tonal language. Each Vietnamese syllable can be considered as a combination of initial, final and tone components. These components include an initial sound, a medial sound, a nucleus sound, a final sound and a tone [7]. The combinations of these sounds and tones are structured in Table I. In total Vietnamese has 21 initial and 155 final components. The total number of pronounceable syllables in Vietnamese is about 19.000 but only about 7.000 syllables (with and without tone) are used in daily language [11]. There are six lexical tones in Vietnamese, which can affect word meaning, i.e. that six different tones applied to the syllables may result in six different words.

TABLE I
Phonological hierarchy of Vietnamese syllables

Tonal syllable				
Initial	Final			Tone
	Medial	Nucleus	Ending	

III. Vietnamese resources

A. Text Corpus

To build a text corpus of Vietnamese words we used the Rapid Language Adaptation Tools [9] to collect text from fifteen websites as listed in Table II, covering main Vietnamese newspaper sources. RLAT enables the user to crawl text from a given webpage with different link depths. The websites were crawled with a link depth of 5 or 10, i.e. we captured the content of the given webpage, then followed all links of that page to crawl the content of the successor pages (link level 2) and so forth until we reached the specified link depth. After

978-1-4244-5479-2/09/$26.00 © 2009 IEEE

collecting the Vietnamese text content of all pages, the text was cleaned and normalized in the following four steps (1) Remove all HTML-Tags and codes, (2) Remove special characters and empty lines, (3) Delete lines with less than 75% tonal words (this step implies identification of Vietnamese language), and (4) Delete lines which appear repeatedly. The first twelve websites of Table II were used to build the language model (see below). The text from the remaining three websites was used to select prompts for recording speech data for the development and evaluation set.

TABLE II
LIST OF ALL 15 VIETNAMESE WEBSITES

	Websites	Link depth
1	www.tintuconline.vn	10
2	www.nhandan.org.vn	10
3	www.tuoitre.org.vn	10
4	www.tinmoi.com.vn	5
5	www.laodong.com.vn	5
6	www.tet.tintuconline.com.vn	5
7	www.anninhthudo.vn	5
8	www.thanhnien.com.vn	5
9	www.baomoi.com	5
10	www.ca.cand.com.vn	5
11	www.vnn.vn	5
12	www.tinthethao.com.vn	5
13	www.thethaovanhoa.vn	5
14	www.vnexpress.net	5
15	www.dantri.com	5

B. Speech Corpus

To develop and evaluate our Vietnamese recognizer, we collected Vietnamese speech data in GlobalPhone style[1] [8], i.e. we asked native speakers of Vietnamese to read prompted sentences of newspaper articles. The resulting corpus consists of 25 hours of speech data spoken by 140 speakers, from the cities of Hanoi and Ho Chi Minh City in Vietnam as well as 20 native speakers living in Karlsruhe, Germany. Each speaker read between 50 and 200 utterances which were collected from the above listed 15 Vietnamese websites. In total the corpus contains 22.112 utterances spoken by 90 male and 70 female speakers. The speech data was recorded in two phases using the GlobalPhone toolkit. In a first phase the speech data G1 was collected from 120 speakers in the cities of Hanoi and Ho Chi Minh. In the second phase we selected utterances from the text corpus in order to cover rare Vietnamese phonemes. This second recording phase was carried out with 20 Vietnamese speakers who study in Karlsruhe (G2). All speech data is recorded with a headset microphone in clean environmental conditions. The data is sampled at 16 kHz with a resolution of 16 bits and stored at PCM encoding. Out of these 160 speakers 140 were assigned for training of the acoustic models. In total the training set consists of 19.596 utterances from the first 12 websites. Ten speakers (4 male and 2 female speakers from North Vietnam, 2 male and 2 female speakers from South Vietnam) were selected for the development set. Another 10 speakers with the same gender and dialect distribution were selected to evaluate the recognition systems. The utterances of the development and evaluation set were chosen from the last three websites (13-15), as listed in Table II. The final Vietnamese portion of the GlobalPhone database is listed in Table III.

TABLE III
VIETNAMESE GLOBALPHONE SPEECH CORPUS

Set	#Speakers		#Utterances	Duration
	Male	Female		
Training	78	62	19596	22h 15min
Development	6	4	1291	1h 40min
Evaluation	6	4	1225	1h 30min
Total	90	70	22112	25h 25min

C. Language Model

We built two statistical n-gram language models using the SRI language model toolkit. Initially, we trained a 3-gram language model on the cleaned and normalized text data from the first 8 websites listed in Table II. We refer to this language model as LM-1. Furthermore, we built a language model LM-2 on an increased text corpus to improve the performance of the recognition system. LM-2 is based on the complete set of the 12 websites used in training, enriched by additional vocabulary from the development set, but does not include any text data from the remaining websites (13-15). Table IV gives the characteristics of these language models.

TABLE IV
PERFORMANCE OF LM-1 AND LM-2

	LM-1		LM-2	
Criteria	Dev	Eval	Dev	Eval
# word tokens	34307750		39043284	
# vocabulary	29920		29967	
OOV-Rate (%)	0.02	0.067	0	0.067
Perplexity	344	347	282	277
Coverage (%):				
1-gram	99.98	99.94	100	99.94
2-gram	88.4	91.28	93.4	92.60
3-gram	50.5	50.04	60	54.02

IV. BASELINE RECOGNITION SYSTEM

The sound structure of the Vietnamese language allows to use various basic modeling units, i.e. initial-finals, phonemes, or a mixture of both. Since in this work our aim was to developed a Vietnamese recognizer using a limited amount of 25 hours of speech data, we decided to use phonemes as basic modeling units.

A. Tone Modeling

To account for the tonal structure of Vietnamese we explored two different acoustic modeling schemes in this work.

[1] GlobalPhone is a multilingual speech and text data collection available in 15 languages from the ELRA Catalogue http://catalog.elra.info

In the so-called ”‘Explicit tone modeling”’ scheme all tonal phonemes (vowels, diphthongs, and triphthongs) are modeled with 6 different models, one per tone. For example, the vowel ’a’ is represented by the models ’a1’, ’a2’, ..., ’a6’, where the numerals identify the tones. This scheme results in 238 phonemes, represented by 715 context independent acoustic models (one begin, middle, and end state per phoneme plus silence model). In the so-called ”‘Data-driven tone modeling”’ we used only one model for all tonal variants of a phoneme, i.e. vowel ’a’ is represented by only one model ’a’. However, the information about the tone was added to the dictionary in form of a tone tag. Our speech recognition toolkit allows to use these tags as questions to be asked in the context decision tree when building context dependent acoustic models. This way, the data will decide during model clustering if two tones have a similar impact on the basic phoneme. If so, the two tonal variants of that basic phoneme would share one common model. In case the tone is distinctive (of that phoneme and/or its context), the question about the tone may result in a decision tree split, such that different tonal variants of the same basic phonemes would end up being represented by different models. For context dependent acoustic modeling we stopped the decision tree splitting process at 2500 quintphones for both schemes, the explicit and the data-driven tone modeling. Table V describes the phoneme set and the relevant characteristics of the two different tone modeling schemes as used in the experiments. While the number of basic model units is quite different, the number of context dependent models was controlled to be the same for both schemes for better comparison. After context clustering, a merge&split training was applied that selects the number of Gaussians according to the amount of data. Please note that the ”‘Explicit tone modeling”’ uses about 16% less Gaussians than the ”‘Data-driven tone modeling”’. This results from the fact that many tonal variants, particularly diphthongs and triphthongs are very rare and are thus modeled with a small number of Gaussians.

TABLE V
PHONEME SET AND MODEL SIZE

	Explicit tone modeling	Data-driven tone modeling
# Consonants	22	22
# Vowels	66	11
# Diphthongs	126	21
# Triphthongs	24	4
$\sum$ Phonemes	238	58
# CI Acoustic Models	715	175
# CD Acoustic Models	2500	2500
# Gaussians (Merge&Split)	111421	130263

We did an analysis of the phoneme frequencies in the speech corpus for the ”‘Data-driven tone modeling”’ system on all training (G1 and G2) and development data. The results are presented in Figure 1 and indicate that despite sharing across tones, some of the phonemes are still extremely rare in the speech data corpus.

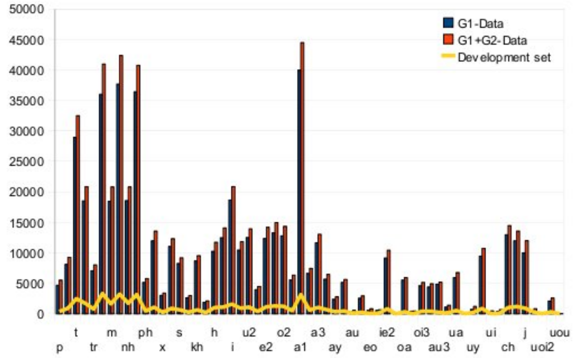

Fig. 1. Phoneme analysis on Vietnamese training and development data

B. Bootstrapping with RLAT System

To rapidly build a baseline recognizer for Vietnamese language we applied our Rapid Language Adaptation Toolkit (RLAT) [9] for bootstrapping the Vietnamese system using a multilingual phone inventory. This phone inventory MM7 was trained from seven randomly selected GlobalPhone languages (Chinese, Croatian, German, English, Spanish, Japanese and Turkish) [1]. To bootstrap the Vietnamese system, the Vietnamese phoneme models were initialized from the closest matches of the MM7 inventory derived by an IPA-based phone mapping. We adopted the GlobalPhone style preprocessing and used the selected MM7 models as seed models to produce initial state alignments for the Vietnamese G1 speech data. The preprocessing consists of feature extraction applying a Hamming window of 16ms length with a window overlap of 10ms. Each feature vector has 43 dimensions containing 13 Melscale Frequency Ceptral Coefficients (MFCC), their first and second derivatives, zero crossing rate, power, and delta power. A Linear Discriminant Analysis transformation is computed to reduce the feature vector size to 32 dimensions. The acoustic model uses a semi-continuous 3-state left-to-right HMM. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. The language model (LM-1) and the pronunciation dictionary are based on monosyllable words. The latter was built by grapheme-to-phoneme mapping, which fortunately is very straight-forward in Vietnamese. Table VI shows the Word Error Rate (WER) performance of the resulting baseline Vietnamese recognizer on the development set after k-means clustering and 4 iterations of Viterbi training based on the initial state alignments produced by RLAT.

TABLE VI
WER AFTER RAPID BOOTSTRAPPING USING RLAT

Systems	Dev-Set
Explicit tone modeling	30.2%
Data-driven tone modeling	27.3%

After rapid bootstrapping we improved the system using Janus Recognition Toolkit (JRTk). Starting from the baseline system we produced improved state alignments and then changed the feature extraction scheme to 143 dimensional features by stacking 11 adjacent frames of 13 coefficient MFCC frames [3]. An LDA transformation is used to reduce the feature vector size to 42. We reiterated k-means clustering and performed 4 iterations of Viterbi training using merge&split training to further improve the system. The results are shown in Table VII. Since the new feature set gave significant improvements we proceeded with this feature extraction scheme for the remainder of the experiments.

TABLE VII
WER WITH NEW FEATURE SET AFER RETRAINING

Systems	Dev-Set
Explicit tone modeling	23.3%
Data-driven tone modeling	24.4%

V. Language Specific System Optimization

In this section we describe the system optimization by introducing methods to tailor the recognizer toward the peculiarities of the Vietnamese language. First, we combined parts of the monosyllabic words to multisyllabic words to enlarge the history of the language model and to improve the modeling of context dependent acoustic model. Due to the restriction of crossword models to one phoneme into the neighbouring word, the lenght of word units has a strong impact on the context dependent modeling. Second, to make the system more robust to noise effects, we adopted a signal adaptation step. Third, to take the described tonal characteristics of Vietnamese into consideration, we used a pitch trajectory extraction method to directly model tones. Finally, we improved the language model by increasing the corpus size and extended the training corpus by adding G2 speech data.

A. Multisyllabic Words

To increase the history in the language model and the context width in the acoustic modeling we combined monosyllable words to multisyllable words by concatenating syllables. For example, the multisyllable word “sinh1_vien1” (student) was merged from “sinh1“ and ”vien1“. For that we had to overcome two problems. First, we had to find suitable multisyllables. To solve this problem we used an open source dictionary from University of Leipzig [5]. It contains about 23.000 bisyllable Vietnamese words and about 6.500 monosyllable words. The second problem was to figure out which syllables should be concatenated. Three methods have been described in the literature: apply statistical information, linguistic information, and a hybrid of both. In this work we rely on the statistical method. Using crawled text data we calculated the frequencies of all bisyllable words adopted from [5]. For each sentence in the text corpus, we searched syllable by syllable for multisyllabic words 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. Using the new text corpus we created a new language model with RLAT. We then concatenated the corresponding syallables in the transcription of the audio data and retrained the acoustic model applying the same parameters as for the above system. The resulting improvements are listed in Table VIII. Apparently, both Vietnamese recognition systems benefit from the multisyallabic approach, with larger gains for the Data-driven tone modeling system. This is probably due to the fact that the larger context helps to improve the clustering of tonal variants.

TABLE VIII
WER FOR MULTISYALLABIC RECOGNITION SYSTEM

Systems	Dev-Set
Explicit tone modeling	21.7%
Data-driven tone modeling	19.5%

B. Signal Adaptation

The Vietnamese GlobalPhone data was recorded under different environmental conditions. To collect data from as many speakers as possible, we recorded sessions in small offices, but also in schools, restaurants, and hospitals. To reduce the affect of background noises, we applied the maximum likelihood signal adaptation to normalize the data on the signal level in the training and development set [4]. In training the speech signal was normalized to reduce the variation of parameters resulting in an acoustic model that is more robust against variation of noise and channel. In the recognition phase the signal was transformed to maximize the conditional probability $p(x|\lambda)$ where x is the signal and λ is the acoustic model. By using signal adaptation the WER of both systems was reduced by about 10% relative (see Table IX).

TABLE IX
WER AFTER SIGNAL ADAPTATION

Systems	Dev-Set
Explicit tone modeling	18.6%
Data-driven tone modeling	17.2%

C. Pitch Extraction

Since Vietnamese is a tonal language, pitch information is an important aspect to improve recognition performance. In this paper we explored three methods according to Schubert [6] to extract pitch information, (1) using cepstrum, (2) using the root cepstrum, and (3) apply autocorrelation. We computed the Cepstrum (autocorrelation or rootcepstrum) with a window length of 40ms and detected the position of the maximum of all cepstral coefficients starting with the 30th coefficient. Furthermore, we considered the position of the three left and right neighbors, and their first and second derivatives. This resulted in 21 additional coefficients (1 maximum, 3 left neighbor, 3 right neigbor plus the first and second order derivatives). These 21 coefficients were added to the original 143 dimensional feature vector. With an LDA transformation we finally reduced this set to 42 dimensions, to keep the number of features the same as in the baseline systems. The results in Table X indicate that extracting pitch information and integrating this information into the recognizer improves the recognition performance on Vietnamese speech. Also, it shows that using the cepstrum achieves the best results in our experiments. Therefore, we applied this method in the remainder of our experiments.

TABLE X
WER AFTER INTEGRATION OF EXPLICIT PITCH INFORMATION

Systems	Explicit tone modeling	Data-driven tone modeling
Cepstrum	17.0%	16.3%
RootCepstrum	17.6%	16.9%
Autocorrelation	17.7%	17.3%

D. Language Model Improvement

So far, all experiments were performed with the language model LM-1. As reported in section III.C, LM-1 was built on the websites 1-8 (see table II), resulting in a low 3-gram coverage, a rather high perplexity of about 344, and an Out-Of Vocabulary rate (OOV) of 0.02% on the development set. In order to improve the language model, we increased the text corpus by additionally crawling the websites 9-12 and also added the vocabulary of the development set. As shown in Table XI, the resulting language model LM-2 gave about 10% relative performance improvement on the development set.

TABLE XI
WER WITH IMPROVED LANGUAGE MODEL LM-2

Systems	Dev-Set
Explicit tone modeling	15.0%
Data-driven tone modeling	14.5%

E. Phoneme Coverage

As shown in Figure 1 some of the phonemes occur very rarely in the Vietnamese G1 training corpus. An error analysis of the best performing system revealed that many of the word errors were a direct consequence of these poorly modeled rare phonemes. To further improve the system performance we considered two options: either collapsing the phoneme set by subsuming the rare phonemes under their closest match and splitting diphthongs and triphthongs into their respective monophthong parts or perform a selective collection of additional speech data in order to increase the occurencies of rare phonemes. Collapsing the phoneme set would have further increased confusability, also splitting up diphthongs and triphtongs overestimates the phoneme duration. We therefore decided to collect an additional 2 hours of speech data (G2) from Vietnamese students who study at the University of Karlsruhe. The training data G1 was then extended by G2 and a new system with same number of model parameters as before was trained. The results in Table XII demonstrate that increasing the occurence of rare phonemes by selectively collecting additional speech data gave a significant performance gain. However, we did not investigate if the gain is solely based on the better phoneme coverage or is a result from using 10% more training data. It can also be observed that the former performance gap between Explicit tone modeling and Data-driven tone modeling now almost nivilizes, indicating that the data-driven method was better equipped to deal with rare phonemes due to the model sharing scheme.

TABLE XII
WER WITH IMPROVED PHONEME COVERAGE

Systems	Dev-Set
Explicit tone modeling	13.6%
Data-driven tone modeling	13.5%

F. Language Model Tuning

Finally, we retuned the language model weights and word insertion penalties by rescoring the word lattices on the development set. This gave another improvement of about 8% relative on WER. Since we optimzed the parameters on the development set, we finally tested our system on the unseen evaluation set. Table XIII shows the results for both test sets, indicating that the performance is rather stable across the sets.

TABLE XIII
WER AFTER LANGUAGE MODEL TUNING

Systems	Dev-Set	Eval-Set
Explicit tone modeling	12.8%	12.2%
Data-driven tone modeling	12.6%	11.7%

VI. DIALECTAL VARIANTS

The data collection took place in both, North and South Vietnam. The two regional dialects differ in their sound systems, but also in vocabulary, and grammar, e.g. the syllable-initials *ch* and *tr* are pronounced the same in the Northern dialect, while kept distinct in the South. Also, the North pronounces *d*, *gi*, and *r* the same, while the South keeps *r* distinct. In Central and Southern varieties the palatals *ch* and *nh* at the end of syllables are no longer distinct from the alveolars *t* and *n*. Additional differences in our data collection stem from the fact that we collected in more noisy places in South Vietnam. To cope with the regional dialects we built a dialect dependent recognition systems by encoding the dialect as an additional tag in the pronunciation dictionary. In the system 'AllW' we doubled the entries of words, and marked one variant with tag 'N' for North and one variante with 'S' for South. The speech from Northern Vietnamese speakers would then be trained using the 'N' variants of words, and the Southern Vietnamese speakers would be trained using 'S'. During the clustering procedure the data decide if Northern and Southern variants of phoneme models are similar enough to share one model. The results of this system (AllW) tested on all speakers of the development set is about 4% absolute WER worse compared to the dialect independent recognizer. The hypotheses revealed that most of the words spoken by Northern test speakers were recognized to be the Northern variant of this word. However, only 20% of the words spoken by Southern speakers were recognized to be Southern variants.

We investigated a second scheme where we marked only those words in the database, which were spoken differently by the South Vietnamese and used only Tag S for these critical words in dictionary. We refer to this system as 'CriW'. The results in Table XIV show that system CriW significantly outperforms system AllW. However, both systems are worse compared to the dialect independent versions (Data-Driven tone modelling before LM tuning), which gave 13.5% on all speakers, with 9.4% WER for Northern speakers and 19.3% WER for Southern speakers. So, while system AllW benefits from improvements on Northern speakers, it significantly hurts the performance on Southern speakers. Table XIV illustrates that the recognition on Northern speakers is drastically better than on the Southern speakers. If this is a consequence of noisy conditions or of poorly modeled dialectal variation will be subject to further investigations.

TABLE XIV
WER FOR DIALECT SYSTEMS FOR "EXPLICIT TONE MODELING" (ETM) AND "DATA-DRIVEN TONE MODELING" (DDTM) SYSTEM

WER	AllW		CriW	
	ETM	DDTM	ETM	DDTM
Northern speakers	9.4%	11.0%	9.6%	9.2%
Southern speakers	28.5%	28.9%	19.7%	19.7%
All speakers	17.3%	18.5%	13.8%	13.6%

VII. CONCLUSION

In this paper we described the development of a Vietnamese speech recognition system for large vocabulary. For this purpose we collected about 25 hours of speech data from 160 Vietnamese speakers reading newspaper articles. For language modeling we collected a text corpus of roughly 40 Mio words. After a rapid bootstrapping based on our Rapid Language Adaptation Toolkit applying a multilingual phone inventory, we improved the performance by carefully investigating the peculiarities of the Vietnamese language. In particular, we implemented and compared different tone modeling schemes and extended the feature set by pitch extraction. To address the monosyllabic structure of Vietnamese, we created multiwords and thus increased the reach of the language model and acoustic model. Furthermore, we addressed the issue of dialectal variations between South and North Vietnam by creating dialect dependent pronunciations and including dialect in the context decision tree of the recognizer. The initial recognition performance of 28% Word Error Rate was improved to 12.6% on the development set and 11.7% on the evaluation set. The impact of the various optimization steps are summarized in Table XV and illustrated in Figure 2. In the future we plan to complement the corpus by a collection of speech data from Central Vietnam and further investigate the impact of dialectal variations. Furthermore, we plan to explore other methods for Vietnamese text segmentation.

TABLE XV
SYSTEM OPTIMIZATION

System	Explicit tone modeling	Data-driven tone modeling
Bootstrapping	30.2%	27.3%
New Feature Set	23.4%	24.4%
Multisyllabic Words	21.7%	19.3%
Signal Adaptation	18.6%	17.2%
Pitch Information	17.0%	16.3%
LM Improvements	15.0%	14.5%
Phoneme Coverage	13.6%	13.5%
LM Tuning	12.8%	12.6%

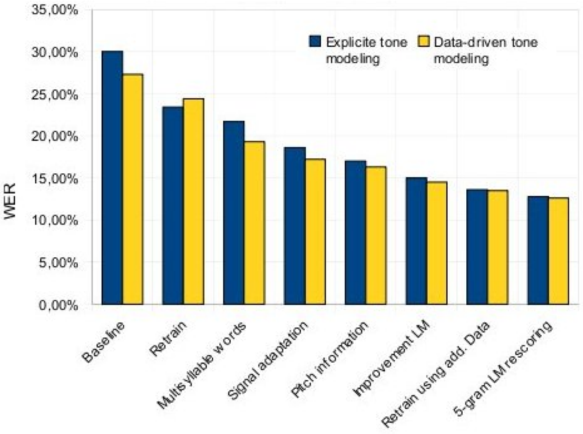

Fig. 2. WER Performance Evolution on Development Set

ACKNOWLEDGMENT

The authors are grateful to all friends and relatives in Hanoi and Ho Chi Minh City in Vietnam, as well as in Karlsruhe, Germany for donating their time, effort, and speech for the Vietnamese database collection. We also would like to thank Tim Schlippe, Emilian Stoimenov, Yik-Cheung (Wilson) Tam, and Micheal Wand for their support and useful discussions.

REFERENCES

[1] Tanja Schultz and Alex Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. In Speech Communication August 2001, Volume 35, Issue 1-2, pp 31-51.

[2] Sinaporn Suebvisai, Paisarn Charoenpornsawat, Alan W Black, Monika Woszczyna, and Tanja Schultz. Thai Automatic Speech Recogntion. In: Proc ICASSP 2005, Philadelphia, PA 2005.

[3] Hua Yu and Alex Waibel. Streamlining the front end of speech recognizer. In: Proc ICSLP 2000.

[4] Ivica Rogina. Optimization of Parameters for Dictation with unlimited Vocabulary. In: Dissertation, University Karlsruhe, Germany 1997.

[5] Leipzig Vietnamese Pronunciation Dictionary. In: http://www.informatik.uni-leipzig-de/d̃uc/Dict/install.html.

[6] Kjell Schubert. Pitch tracking and his application on speech recognition. In: Diploma Thesis at University of Kalsruhe(TH), Germany, 1998.

[7] Nguyen Hong Quang, Pascal Nocera, Eric Castelli, Trinh Van Loan. A Novel Approach in Continuous Speech Recognition for Vietnamese, an isolating tonal language. In: SLTU, Hanoi, Vietnam, 2008.

[8] Tanja Schultz. GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University. In: Proc. ICSLP Denver, CO, 2002.

[9] Tanja Schultz and Alan Black. Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing. In: Proc ICASSP Las Vegas, NV 2008.

[10] A M Noll. Short-Time Spectrum and "Cepstrum" Techniques for Vocal-Pitch Detection. In: J. Acoust. Soc. Am. Volume 36, Issue 2, pp. 296-302, 1964.

[11] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong, John-Paul Hosom. Vietnamese Large Vocabulary Continuous Speech Recognition. In: Interspeech, Lisbon Portugal, 2005.