OPTIMIZATION ON VIETNAMESE LARGE VOCABULARY SPEECH RECOGNITION

Ngoc Thang Vu, Tanja Schultz

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

ABSTRACT

This paper summarizes our latest efforts toward a large vocabulary speech recognition system for Vietnamese. We describe the Vietnamese text and speech database which we collected as part of our GlobalPhone corpus. Based on these data we improve our initial Vietnamese recognition system [1] by applying various state-of-the art techniques such as semi-tied covariance and discriminative training. Furthermore, we achieve significant improvements by building two systems based on different tone modeling approaches and then apply system cross-adaptation and confusion networks combination. The best Vietnamese speech recognition system employs a 3-pass decoding strategy and achieves a syllable-based error rate of 7.9% on read newspaper speech. In addition, we perform initial experiments on the Voice of Vietnam (VOV) speech corpus [2] and achieve a syllable error rate of 16.5%.

Index Terms— Vietnamese speech recognition, data collection, discriminative training, system combination

1. INTRODUCTION

The performance of speech and language processing technologies has improved dramatically and an increasing number of systems are being deployed in a large variety of applications. To date, most efforts were focused on a very small number of languages spoken by a large number of speakers in countries of great economic potential, and a population with immediate information technology needs. With more than 6900 languages in the world and the need to support multiple input and output languages, the most important challenge today is to port or adapt speech processing systems to unsupported languages rapidly and at reasonable costs. Despite the fact that the Vietnamese language is spoken by more than 80 Million people and thus is listed among the top-25 languages, there is a surprisingly small number of groups investigating Vietnamese speech and language processing technologies and applications, with notable exceptions like IOIT [2] and MICA [3].

Last year we started to applying our Rapid Language Adaptation Tools (RLAT) [4] to Vietnamese. In [1] we reported on our development and optimization of a Vietnamese large vocabulary speech recognition system and described particular characteristics of the Vietnamese language, such as the monosyllabic structure and tonality of the sound system. Our best system achieved a syllable error rate (SyllER) of 12.6 % on the development and 11.7% on the evaluation set. However, this initial system did not employ the full range of state-of-the-art techniques, which have shown to be very effective for high-resource languages. In this paper we apply these techniques to our initial Vietnamese system and study to what extend the reported performance improvements on languages like English and Chinese apply to Vietnamese. Among the state-of-the-art techniques we applied are semi-tied covariances [5], discriminative training [6], system cross adaptation, and confusion network combination [7].

The paper is organized as follows. In Section II we describe our Vietnamese resources, which consist of a audio data and corresponding transcriptions in the newspaper domain, and a large text corpus harvested from the internet on the same domain. Section III introduces our baseline recognition system which was presented in [1]. In Section IV we give a detailed description of the optimization steps and report recognition results on the development and evaluation set. The study is concluded in Section VI with a summary and an outlook to future steps.

2. VIETNAMESE LANGUAGE PECULIARITIES

Vietnamese is a language with very interesting characteristics, three of which are particularly challenging for automatic speech recognition. The first peculiarity is the monosyllabic nature of Vietnamese. For example the sentence ”Xin chao Viet Nam“ (in English: hello Vietnam) contains 4 word units, each consisting of a single syllable. This monosyllabic nature poses two problems to speech recognition, i.e. due to the shortness, the word units are acoustically confusable and the short units limit the language model history. In [1] we compensated the restricted language model history by concatenating monosyllabic words to multisyllabic words. After concatenation, the example sentence from above looks like ”Xin_chao Viet_Nam“. The sentence has now 2 multisyllabic words. Multisyllabic words achieve significant improvements ranging from 10% to 20% relative, depending on the tone modeling approaches.

The second peculiarity of the Vietnamese language is the tonality of the sound system. Vietnamese has six different

tones, which can discriminate the meaning of words. So, it is advisable to use tone information in the acoustic model. In [1] we extracted pitch information using the Cepstrum and gained about 6% to 9% relative improvement depending on the tone modeling approaches.

The third important characteristic of Vietnamese results from the large amount of diphthongs and triphthongs in the phoneme set. In total, Vietnamese has 22 consonants, 11 vowels, 21 diphthongs and 3 triphthongs. So, compared to languages like English or French, the number of diphthongs and triphthongs is pretty high. In addition to the large number, some of these phonemes are very rare, and thus may lead to poorly estimated acoustic models. While it is possible to collapse the phone set by subsuming the rare phonemes under their closest match, or by splitting the rare diphthongs and triphthongs into their respective monophthongs parts, both approaches have disadvantages. Collapsing the phoneme set results in an increased confusability, and splitting up diphthongs and triphthongs overestimates the phoneme duration. Therefore, we decided in our study to collect additional data to cover rare diphthongs and triphthongs. As reported in [1] we achieved about 8% relative improvement. These gains suggest that for Vietnamese speech recognition care needs to be taken to collect a corpus such that it covers all phonemes.

3. VIETNAMESE LANGUAGE RESOURCES

The development of a state-of-the-art speech recognition system starts with collecting speech data and corresponding transcriptions, as well as written text resources for vocabulary selection and language modeling. Data collection is an extremely time and cost consuming task but its careful excecution is crucial to the performance of the final system. We applied our Rapid Language Adaptation Tools (RLAT) [4], which allow us to collect massiv amounts of text data from the web and to record speech data over the Internet using a web-based recorder. In the following subsections we describe the collected corpus for Vietnamese language that was collected in 2009 as part of our GlobalPhone project [8].

3.1. Text Corpus

For the text corpus of Vietnamese words we used RLAT to collect text from fifteen different websites, 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 collecting the Vietnamese text content of all pages, the text was cleaned and normalized with four different steps: (1) Remove all HTML-Tags and codes, (2) Remove special characters and empty lines, (3) Delete lines with less than 75% tonal words (identification of Vietnamese language) and (4) Delete line which appear repeatedly. The first twelve websites of Table 1 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. In total we collected roughly 40 Million Vietnamese word tokens (see 4 below).

Table 1. List of all 15 Vietnamese websites

Websites	Link depth
www.tintuconline.vn	10
www.nhandan.org.vn	10
www.tuoitre.org.vn	10
www.tinmoi.com.vn	5
www.laodong.com.vn	5
www.tet.tintuconline.com.vn	5
www.anninhthudo.vn	5
www.thanhnien.com.vn	5
www.baomoi.com	5
www.ca.cand.com.vn	5
www.vnn.vn	5
www.tinthethao.com.vn	5
www.thethaovanhoa.vn	5
www.vnexpress.net	5
www.dantri.com	5

3.2. Speech Corpus

3.2.1. GlobalPhone Data

To collect Vietnamese speech data in a very short time, the author spent one month in Vietname and recruited friends and relatives to donate their voice for research. The web-based recording tool turned out to be difficult as many sites in Vietname did not provide Internet connection, so we used an offline version of the same recording tools. In order to control the quality of recordings and to avoid the amount of transcription work, we collected Vietnamese speech data in GlobalPhone style [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 native 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 different Vietnamese websites. In total the corpus contains 22.112 utterances spoken by 90 male and 70 female speakers. All speech data was 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. The Vietnamese portion of the GlobalPhone database is listed in Table 2.

Table 2. 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

3.2.2. Voice of Vietnam Data

The Voice of Vietnam (VOV) speech corpus was collected in 2005 by IOIT and kindly provided to us for research purposes [2]. The VOV data is a collection of story reading, VOV mailbag, news report and colloqiums from the radio program "The Voice of Vietnam". The database consists of are 22549 audio files with transcriptions from 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 [2]. The VOV corpus covers all Vietnamese phonemes and most Vietnamese syllables. The data is provided in wav format, using a sampling rate of 16kHz and A/D conversion precision of 16 bits. We splitted the VOV data in a training and testing part. Table 3 shows the relevant information about the VOV corpus for the training and the test set.

Table 3. The Voice of Vietnam Speech corpus

Set	#Utterances	Duration
Training	20990	19h 31min
Testing	1459	1h 18min
Total	22549	20h 49min

3.3. Language Model

Based on the crawled text corpus (see above), we built a statistical n-gram language model using the SRI language model toolkit [9]. We trained a 5-gram language model on the cleaned and normalized text data from the 12 first websites listed in Table 1. Table 4 gives the characteristics of the language models calculated on the GlobalPhone development set, evaluation set, and VOV test set.

3.4. Pronounciation Dictionary

Next to the speech and text data, the pronounciation dictionary is a very important part of an automatic speech recognition system. The dictionary guides the decoder and ensures proper training alignment. We used the RLAT tools to generate the dictionary. In RLAT an interactiv rule-based lexlearner is implemented which enable the user to learn pronunciation rules by providing initial letter-to-sound mappings and interactively confirming or correcting pronunciation examples as proposed by the lexlearner. We took the RLAT dictionary

Table 4. Performance of LM in development and evaluation set

Criteria	GP-Dev	GP-Eval	VOV-Test
# word tokens	39043284		
# vocabulary	29967		
OOV-Rate (%)	0	0.067	0.11
Perplexity	282	277	392
Coverage (%):			
1-gram	100	99.94	99.89
2-gram	93.4	92.60	92.99
3-gram	60	54.02	54.84
4-gram	32.6	24.2	20.01
5-gram	21.3	12.1	5.8

and performed some manual corrections. More particularly, we wanted to model the impact of dialectal variations by using pronunciation variants. The data were intentionally collected in the North and South of Vietnam and many words are spoken different between the Northern and Southern dialect. Table 5 shows some examples from our pronunciation dictionary applying pronunciation variants.

Table 5. Pronunciation dictionary with different variants for Northern and Southern dialect in Vietnamese

Words	Pronunciation
xin_chao	{x i11 n ch ao2}
vo	{v o36}
vo(1)	{j o36}
ra	{r a11}
ra(1)	{d1 a11}

4. BASELINE RECOGNITION

To model the tonal structure of Vietnamese we explored two different acoustic modeling schemes. In the so-called "'Explicite tone modeling"' (ETM) 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. In the so-called "'Data-driven tone modeling"' (DDTM) 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. The Janus Recognition Toolkit (JRTk) [10] allows using 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 explicite and the data-driven tone modeling. Table 6 describes the phoneme set and the relevant characteristics of the two different tone modeling schemes as used in the experiments reported below. While the number of basic model units is quite different for the two modeling schemes, 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, which selects the number of Gaussians according to the amount of data. Please note that the "'Explicite tone modeling"' uses about 16% fewer Gaussians than the "'Data-driven tone modeling"'. This is a result from the fact that many tonal variants, particularly diphthongs and triphthongs are very rare and are thus modeled with a small number of Gaussians. The preprocessing

Table 6. Phoneme set and model size

	Explicite 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

consists of feature extraction applying a Hamming window of 16ms length with a window overlap of 10ms. Each feature vector has 164 dimensions containing two main parts. The first part has 143 dimensions which were extracted by stacking 11 adjacent frames of 13 coefficient MFCC frames. The second part describes the tone information. We computed the Cepstrum 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 neighbors, 3 right neighbors plus the first and second order derivatives). With an LDA transformation we finally reduced this set to 42 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 and the pronunciation dictionary are based on bisyllable words. Table 7 shows the Syllabic Error Rate (SyllER) performance of the resulting baseline Vietnamese recognizer on the development set after merge-and-split training and 6 iterations of Viterbi training.

Table 7. SyllER of the baseline system on development set

Systems	**GP Dev-Set**
Explicite tone modeling	12.8%
Data-driven tone modeling	12.6%

5. SYSTEM OPTIMIZATION

In this section we describe the steps and techniques taken to optimize the performance of the recognition system. As a first step we applied semi-tied covariances [5] to make the system more robust, for example if training data and test data were recorded in different environments. Second, we ran discriminative training [6] and describe the effect on our speech recognizer. Third, we used cross-adaptation, one of the multi-pass decoding strategies, to combine the advantages of the two different tone modeling approaches, which were implemented as described above. Finally, to minimize the syllabic error rate we used confusion network combination [7] which allows to extract better hypothesis from a combination of two or more systems.

5.1. Semi-tied Covariance Matrices

There is normally a simple choice made in form of the covariance matrix to be used with continuous-density HMMs. Either a diagonal covariance matrix is used, with the underlying assumption that elements of the feature vector are independent, or a full or block-diagonal matrix is used, where all or some of the correlations are explicitly modeled. Unfortunately, full or block-diagonal covariance matrices come with a dramatic increase in the number of parameters per Gaussian component, and thus limiting the number of components which may be estimated robustly. Semi-tied covariance matrices (STC) [5] are a form of covariance matrix which allows a few full covariance matrices to be shared over many distributions, whereas each distribution contains its own diagonal covariance matrix. Furthermore, this technique fits well within the standard maximum-likelihood criterion used for HMM training. Table 8 shows the SyllER performance of the Vietnamese recognizer on the development set after applying semi-tied covariance matrices.

Table 8. SyllER after using Semi-tied Covariance Matrices

Systems	**Dev-Set**
Explicite tone modeling	11.9%
Data-driven tone modeling	11.8%

After this step 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 4% relative in SyllER. Table 9 shows our results on the development set.

Table 9. SyllER after Language Model Retuning

Systems	Dev-Set
Explicite tone modeling	11.7%
Data-driven tone modeling	11.4%

5.2. Discriminative training (DT)

Discriminative training is an essential technique that consistently leads to significant improvements in speech recognition accuracy. Maximum mutual information estimation (MMIE) [11] and boosted MMIE [6] are common techniques for discriminative training. We applied this technique to our Vietnamese speech recognizer system. Starting with the speaker-independent model using maximum likelihood estimation, we decoded the complete set of training utterances in order to generate word lattices.

MMIE aims at maximizing the posterior probability of a reference compared to the competing hypotheses in a word lattice. The objective function of MMIE is:

$$F_{MMI}(\lambda) = \sum_{r=1}^{R} log \frac{P_\lambda(X_r|M_{s_r})P(s_r)}{\sum_s P_\lambda(X_r|M_s)P(s)}$$

where λ represents model parameters to be optimized; X_r is the r-th training utterance; s_r is the reference and M_s represents the corresponding HMM state sequence of sentence s. Maximizing F_{MMI} improves the posterior probability of the reference in the lattice.
Intuitively, some paths may contain more error than other parts in a word lattice. Boosted MMIE boosts the importance of competitors that make large errors and aims to improve the confusable parts. Table 10 shows our results on the development set after applying the discriminative training. So far, we do not have a good explanation why the gains are smaller than expected.

Table 10. SyllER after applying discriminative training

Systems	Dev-Set
Explicite tone modeling	11.56%
Data-driven tone modeling	11.15%

5.3. Multi-pass decoding: Cross Adaptation

State-of-the-art speech recognition systems commonly use multi-pass decoding with an adaptation of the acoustic model between passes. Adaptation aims at better fitting the system to the speakers and/or acoustic environments found in the test data. The two most popular adaptation methods, which can be found in many systems, are Maxmum Likelihood Linear Regression MLLR, a model transformation and Feature Space Adaptation FSA, a feature transformation. Adaptation is performed in an unsupervised manner, so that the hypotheses obtained from the previous decoding pass are taken as the necessary reference for adaptation. Generally, the word error rates of the hypotheses obtained from the adapted systems are lower than without adaptation. This sequences of adaptation and decoding make it possible to incrementally improve the system, but not always lead to significant improvements. Often, after two or three stages of adapting a system on its own output, no more gains can be obtained. This problem can be solved by adapting a system on the output of a different system, a process called cross-system adaption. In this paper we developed distinct systems with two different approaches for tone modeling. Therefore, it is possible to apply cross-system adaptation. Furthermore, for each tone modeling approach we had two different systems: a Speaker Independent (SI) and a Speaker Adaptive (SA) using FSA and MLLR. So we experimented with various possible system combination to find the best performing decoding strategy. As first pass we always apply the SI system. The second and third pass systems are speaker adaptive system. Furthermore, the third pass system could apply the discriminative training. Table 11 shows the results on the development set after applying the various options of cross-system adaptation.

Table 11. SyllER after using Cross Adaptation

Systems	Dev-Set
ETM x DDTM x ETM (S1)	8.7%
ETM x DDT x ETM+DT (S2)	8.4%
ETM x ETM x DDTM (S3)	8.6%
ETM x ETM x DDTM+DT (S4)	8.6%
DDTM x ETM x DDTM (S5)	8.7%
DDTM x ETM x DDTM+DT (S6)	8.6%
DDTM x DDTM x ETM (S7)	8.7%
DDTM x DDTM x ETM+DT (S8)	8.5%

5.4. Confusion Network Combination

After applying the cross adaptation techniques we got different word lattices which contain alternative hypotheses. Consequently, we applied the confusion network combination technique [7] to combine these lattices and subsequently extract the best hypothesis. We experimented with different lattice combinations. The best combination gave 0.2% absolute improvement. Table 12 shows the all results on the development set after applying confusion network combination.

5.5. Decoding strategy

After the optimization steps on the development set we obtained the best decoding strategy. Two parallel systems decode the audio data and write the word lattices. After that we used confusion networks (CN) to combine these lattices and extract the best hypothesis. The first system (S1) contains 3

Table 12. SyllER after using Confusion Network Combination

Systems	Dev-Set
S2 x S6	8.2%
S2 x S8	8.4%
S2 x S4	8.3%
S6 x S8	8.3%
S4 x S6	8.5%
S2 x S4 x S6	8.4%
S2 x S4 x S8	8.4%
S4 x S6 x S8	8.5%
S2 x S4 x S6 x S8	8.3%

passes: ETM-SI, DDTM-SA, and ETM-SA using DT. The second system (S2) contains also 3 passes: DDTM-SI, ETM-SA and DDTM-SA using DT. We tested our system on the unseen evaluation set using this decoding strategy. Table 13 illustrates the results on the evaluation set.

Table 13. SyllER on the evaluation set using the best decoding strategy

	1.Pass	2.Pass	3.Pass	CN
S1	11.4%	8.7%	8.1%	7.9%
S2	10.8%	8.8%	8.2%	7.9%

5.6. Experiments and Optimization on VOV Data

5.6.1. Experiments with VOV data

The VOV corpus was collected from the audio program "Voice of Vietnam". It has substantially different characteristics compared to the GlobalPhone data. As a result the VOV data provide us with a good test case to explore how well our Vietnamese speech recognizer generates. The first experiment applied the "Explicite-tone modeling system" (ETM) to decode the VOV test set and gave 24.1% SyllER. In the second experiment we trained the speech recognition system on the VOV training data and tested on the VOV test data. We used the ETM system to write the initial alignments for the complete VOV training set. We used these initial alignments to train the system. For system training we applied the same parameter settings as we used to train our best GlobalPhone system. The performance on the VOV test set slightly improves to 23.5% but gets drastically worse on the GlobalPhone development set with 33.4% SyllER. According to our analysis, we believe that the reason for the degradation is that the VOV corpus contains only Northern dialect data, while the GlobalPhone data set covers Northern and Southern dialect. The breakdown for dialects shows that the GlobalPhone part with Northern dialect achieved a performance of 19.6% SyllER, while the Southern dialect part significantly dropped in performance to 51.7% SyllER. So, training on Northern-only VOV data significantly harms the performance on the part of GlobalPhone spoken by Southern Vietnamese speakers. In our last experiment we trained the acoustic model with a combination of GlobalPhone and VOV training data. The results are given in Table 14 and show improvements of about 25% relative on the VOV test set, but 5% degradation on the GlobalPhone development set. A subsequent error analyis of these results indicate that the majority of errors stem from the following issues: (1) large number of proper names, sometimes even a sequence of several proper names, (2) interruptions, unfinished utterances (3) Foreign proper names, most particular English, such as Canada, Vovnews and Singapore. In the following section we describe how the language model was trained to better handle proper names and compensate for the above described issues.

Table 14. SyllER on the VOV test set and GP development set using the speaker independent system

Training-Set	VOV Test	GP dev
GP Daten	24.1%	11.9%
VOV Daten	23.5%	33.4%
VOV+GP Daten	17.8%	12.5%

5.6.2. System Optimization on VOV data

In order to adapt our language model to the VOV test set, we used the RLAT system to crawl the VOV mailbag from 22-12-2008 to 22-12-2009 and built a 3-gram language model "VOVmail". Linear interpolation [9] was applied to combine the background and VOVmail language model (LM). The best mixture weight is 0.57 for the background LM and 0.43 for the VOVmail language model. To solve the problem with proper names, we randomly generated 1 million full names and built a 3-gram language model called "FullName". A Vietnamese proper name contains usually three parts: surname, middle name, and firstname. In our work we used the 20 most common surnames, the 35 most common middle names, and 65 of the most common first names and combined them randomly. After that we interpolated the three language models and decoded the VOV test set. Table 15 compares the performance of the baseline language model (background), the interpolation with the VOVmail-based language model (+VOVmails), and the interpolation with the VOVmail data and the automatically generated corpus of full names. The results show that the new language model shows significant perplexity reduction on the VOV test data. Our currently best system gives a SyllER of 16.5% on the VOV test set using the interpolation of all three corpora. This is a gain of 7% relative over the baseline language model.

6. CONCLUSION

In this paper we describes our latest improvements to our Vietnamese speech recognition system for large vocabulary.

Table 15. Optimizing LM on VOV dev set

Criteria	Background	+VOVmails	+FullName
OOV-Rate (%)	0.11	0.04	0.04
Perplexity	392	250.4	245.9
Coverage (%):			
1-gram	99.89	99.96	99.96
2-gram	92.99	94.2	94.26
3-gram	54.84	57.05	57.6
4-gram	20.01		20.01
5-gram	5.8	5.8	5.8

The speech corpus as a part of GlobalPhone was used with 25 hours audio data from 160 Vietnamese speakers reading newspaper articles. Applying our Rapid Language Adaptation Tools, we collected about 40 Mio words from 15 different websites for language model training and prompt selection. We subsequently applied state-of-the-art techniques, such as semi-tied covariance matrices, discriminative training, cross adaptation, and confusion network combination to study the impact on Vietnamese speech recognition and to improve our system. Starting from a baseline system with 12.6 % SyllER, we improved the system to 8.2% on the development set, and reduced the error from 11.7% to 7.9% on the evaluation set. The impact of the various optimization steps and the best decoding strategy are summerized in Table 16 and Table 17. Future steps will include further improvements of tone modeling, language modeling, and a more detailed investigation of the effects of dialects.

Table 16. System Optimization

System (SI)	**Explicite tone modeling**	**Data-driven tone modeling**
Baseline	12.8%	12.6%
Optimal Feature	11.9%	11.8%
LM Tuning	11.7%	11.4%
Discriminative Training	11.56%	11.15%

Table 17. SyllER on development set using the best decoding strategy

Decoding-Pass	**S1**	**S2**
1.Pass	11.7%	11.6%
2.Pass	9.0%	9.0%
3.Pass	8.4%	8.6%
Confusion Network	8.2%	8.2%

7. ACKNOWLEDGMENT

The authors are very grateful to Dr. Luong Chi Mai from IOIT for providing us the VOV speech corpus. We also would like to thank Roger Hsiao for his support and useful discussions.

8. REFERENCES

[1] Ngoc Thang Vu and Tanja Schultz. Vietnamese Large Vocabulary Continuous Speech Recognition. In: ASRU, Italy 2009.

[2] Thang Tat Vu, Dung Tien Nguyen, Mai Chi Luong and John-Paul Hosom. Vietnamese Large Vocabulary Continuous Speech Recognition. In: 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 2005.

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

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

[5] Mark Gales, Semi-tied covariance matrices for hidden Markov models. In: IEEE Transactions Speech and Audio Processing, vol. 7, pp. 272-281, 1999.

[6] Dan Povey, D. Kanevsky, Brian Kingsbury, B. Ramabhadran, George Saon and K. Visweswariah. Boosted MMI for model and feature-space discriminative training. In: Proc. of the IEEE International Conference on Acoustic, Speech and Signal Processing, 2008.

[7] Lidia Mangu, Eric Brill and Andreas Stolcke. Finding Consensus Among Words: Lattice-Based Word Error Minimization. In Proc. of EUROSPEECH'99, Budapest, Hungary.

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

[9] Andreas Stolcke. SRILM - an extensible language modeling toolkit, in Proceedings of ICSLP, 2002.

[10] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal, "The Karlsruhe Verbmobil Speech Recognition Engine," International Conference on Acoustics, Speech, and Signal Processing 1997, ICASSP, Munich; Germany.

[11] V. Valchev, J. Odell, Phil Woodland and Steve Young. MMIE training of large vocabulary speech recognition systems. In: Speech Communication, vol. 22, pp. 303-314, 1997.