Journal of Computer Science and Cybernetics, V.34, N.4 (2018), 365–372
DOI 10.15625/1813-9663/34/4/13181

ACCELERATION IN STATE-OF-THE-ART ASR APPLIED TO A VIETNAMESE TRANSCRIPTION SYSTEM

NHUT M. PHAM[1,a], QUAN H. VU[2]

[1] *University of Science, Ho Chi Minh city, Viet Nam*
[2] *Institute for Computational Science and Technology*
[a] *vhquan@fit.hcmus.edu.vn*

Abstract. This paper presents the adoption of state-of-the-art ASR techniques into Vietnamese. To better assess these techniques, speech corpora in the research community are assembled, and expanded, making a unified evaluation material under the name VN-Corpus. On this corpus, three ASR systems are built using the conventional HMM-GMM recipe, SGMM, and DNN respectively. Experimental results crown DNN with the overall WER of 12.1%. In the best case, DNN even cuts down to 9.7% error rate.

Keywords. Vietnamese automatic speech recognition; Transcription system.

1. INTRODUCTION

Research and findings in Vietnamese automatic Speech Recognition (ASR) have been stagnant for the last few years. Causal factors and catalysts that precipitated the situation include the lack of inspiring approaches, changes in research direction and trend. The most up-to-date Vietnamese ASR engine, deployed in common speech applications [9, 10], makes use of the standard HMM-GMM recipe [3]. This is quite inapt compared to other advanced techniques. After the rise of SGMM [4] and DNN [1] along with their impressive performances, we picked up our pace and resumed the work on Vietnamese ASR, thus making the motivation to push further.

Table 1. Diacritics in Vietnamese

Diacritic	none	/	\	?	~	.
Example	xa	xá	xà	xả	xã	xạ
Meaning	far	bow	snake	release	village	musk

For a brief introduction, Vietnamese is a monosyllable, tonal language. Each word unit is pronounced as a syllable and its meaning depends on the tone. There are about 6596 phonetically distinguishable syllables [2] which comprise of legal combinations between basic

syllables (i.e., syllables without tone) and five tones. Table 1 illustrates the diacritics used for representing tones, including: level tone (denoted by "none"), high-rising tone (/), low-falling tone (\), dipping-rising tone (?), high-rising glottalized tone (˜), and low glottalized tone (.). Although word, a group of one to several syllables, is the smallest syntactically meaningful unit, syllable is the basic pronunciation unit in Vietnamese speech. Hence, using syllable as a basic lexical unit is an ideal choice for Vietnamese ASR.

Earlier works focus on refining the acoustic model [7], domain adjustments [8], and graph twitching [11]. But none has seriously taken into account the nature of tones and their impact on the overall performance. Furthermore, the findings are diverse, each with their own evaluation datasets. The Vietnamese ASR research community really needs a common source of data and an adoption of state-of-the-art techniques. So here, we move on with two parallel but dependent tasks: (1) building a standard Vietnamese speech corpus as a unified evaluation material; and (2) adopting SGMM and DNN into Vietnamese ASR with the attention of tone and acoustic modeling. We also setup and facilitate the conventional HMM-GMM system for comparison purposes.

The rest of this paper is organized as follows. Section 2 presents the unified Vietnamese speech corpus. Section 3 covers our ASR systems and their experimental results. Section 4 concludes the paper.

2. THE VN-CORPUS

Before the establishment of VN-Corpus, experiments of Vietnamese ASR were conducted on several local corpora and recorded data, including:

- The VOH corpus (broadcast news): Consisting of roughly 21 hour speech from 17 speakers (6 males, 11 females) with Southern dialect.
- The VOV corpus (broadcast news): Made up from 18 hour speech of 20 speakers (8 males, 12 females) with Northern dialect.
- The LAB corpus (conversational speech): Composed of 28 hour closed-mic recording sessions from 158 speakers who are students in the university.

2.1. Composing

The construction of VN-Corpus starts off with VOH, VOV, and LAB in hand. So we got 2 categories to fill in: news and conversations.

For the news, we proceed to download video clips from the official national TV channels, including VTV, HTV, and FBNC. Audio streams are extracted from the clips, and then manually segmented and transcribed to remove any non-speech segments such as music, ads, or background noise.

For the conversations, we launched 2 additional recording campaigns to extend the LAB corpus, one in the University of Science, and the other in the School of Dramatic Arts. A total of 103,239 dramatic spoken scripts were chosen for recording. These scripts cover 4951 vocabulary entries, efficiently balance out 93% of the lexical span. Recordings were taken place in a quiet room with closed-mic setting.

All speech data is then sampled to a common format of PCM, 16 KHz, 16 bits, mono.

2.2. Partitioning

After making the speech and their transcriptions ready, we divide them into 3 subsets: the training, the development, and the test set. Details are given in Table 2.

Table 2. The VN-Corpus

		Training set	Development set	Test set	Total
News	VOH	18h	1h	2h	21h
	VOV	15h	1h	2h	18h
	VTV	73h	1h	14h	88h
	FBNC	11h	1h	2h	14h
	HTV	78h	1h	8h	87h
Conversations	LAB	24h	1h	3h	28h
	LAB-expansion	104h	1h	13h	118h
Total		323h	7h	44h	**374h**

These subsets are used to train, fine-tune, and test the ASR systems presented in Section 3. The corpus was also published for academic usage, under the name VN-Corpus.

3. ASR SYSTEMS

3.1. Language modeling

The language model (trigram) was built with the 273M-word text corpus collected from online news and forum threads available on the Internet between 4/2010 - 11/2014. Transcriptions of the training set are also blended in (i.e., interpolation) to make content variation. Abbreviation and numeric expression occurring in the text are then replaced by their written words. The vocabulary contains 5281 words, a combination of words in audio transcriptions and those occurring at least 12 times in the text corpus; thus made an OOV rate of 2.6%.

Table 3. Language model perplexities

	Without Interpolation	With Interpolation
Perplexity	212.6	135.8

To evaluate the language model, 3000 sentences containing 56k tokens are randomly selected from the test set transcription. Table 3 reports perplexities of the language models with and without the joining of audio transcription. It is obvious that the perplexity of the interpolated LM was dramatically reduced, from 212.6 to 135.8, ensuring better performance for the ASR systems.

3.2. Acoustic modeling

Modeling of acoustic data is formerly designed following the Chinese approach [3] in which each syllable is decomposed into initial and final parts. While most of Vietnamese syllables consist of an Initial and a Final, some of them have only the Final. The initial part always corresponds to a consonant. The final part includes main sound plus tone and an optional ending sound. This decomposition results in a total number of 44 monophones. It has two advantages. First, the number of monophones is relatively small. Second, by treating tone as a distinct phone, followed immediately after the main sound, the context-dependent model for tone can be built straightforwardly. It means that the recognition of tones was fully integrated in the system in just one recognition pass. However, distinct representations of tones have brought upon a disadvantage: the deficiency in modeling tonal features (i.e., pitch) across a syllable. Since tones are stressed on the main vowels, separating tone from vowel would degrade the parameterization of tonal vowels.

S → [I] F

F → V [E]

I → b c ch d đ g gh gi h k kh l m n
ng ngh nh p ph qu r s t tr th v x

V → a ă â e ê i o ô ơ u ư y
á ắ ấ é ế í ó ố ớ ú ứ ý
à ằ ầ è ề ì ò ồ ờ ù ừ ỳ
ả ẳ ẩ ẻ ể ỉ ỏ ổ ở ủ ử ỷ
ã ẵ ẫ ẽ ễ ĩ õ ỗ ỡ ũ ữ ỹ
ạ ặ ậ ẹ ệ ị ọ ộ ợ ụ ự ỵ

E → c ch ng nh m n p t

Figure 1. Integrated tone phoneme set

To better model the tonal feature, a modification to the acoustic model is needed, in which tones are integrated into tonal vowels. This results in a new decomposition consisting of 99 monophones including 27 phones for consonants, 12 phones for non-tonal vowels, and 60 phones for tonal vowels as shown in Figure 1. Table 4 gives examples showing the differences between tone representations.

Table 4. Samples of tone sepresentations

	Separated tone	Integrated tone
ngày	ng a \ y	ng à y
ngay	ng a y	ng a y
nghệ	ngh ê .	ngh ệ

Using this decomposition scheme, we worked on 3 different ASR systems. The following

Subsections will take turn to describe them.

3.2.1. Baseline system

The first system is based on the conventional HMM-GMM structure which was introduced in [1]. Composed features, including pitch, 12 MFCCs, energy, their first and second derivatives, are modeled for each of the context-dependent phonemes (triphones). The trained recognizer contains 3861 tied-states with 16 Gaussian mixtures per state distribution.

Table 5. Baseline performances

	WER		
	Broadcast news	Conversations	Overall
Baseline	34%	41.3%	36.6%
Baseline + fMLLR	30.4%	35.2%	32.1%
Baseline + fMLLR + MMI	24.8%	28.9%	26.2%

We also make 2 baseline augmentations: (1) one with additional fMLLR technique [5] as a Speaker Adaptive Training (SAT) recipe, (2) the other using discriminative training with Maximum Mutual Information (MMI) criteria. Performances obtained by these settings are reported in Table 5.

3.2.2. SGMM system

The second system is built following the renowned SGMM technique which was originally formulated under low-resourced conditions [4]. In SGMM, each state distribution is modeled by a mixture of state vectors instead of a GMM as shown in Figure 2. These vectors are indeed projections from a pool of collective Gaussian functions, called by the name Universal Background Model (UBM). Our UBM consists of 800 Gaussian components. An SGMM configuration of 40-dimensional state vectors, and 12 sub-states per state was chosen on the development set.

Table 6. SGMM performances

	WER		
	Broadcast news	Conversations	Overall
SGMM	20.1%	26.7%	22.5%
SGMM + fMLLR	18.5%	25.4%	21%
SGMM + fMLLR + MMI	17.6%	23%	19.6%

Same case with the baseline, SGMM system also got 2 augmentations: fMLLR and MMI.

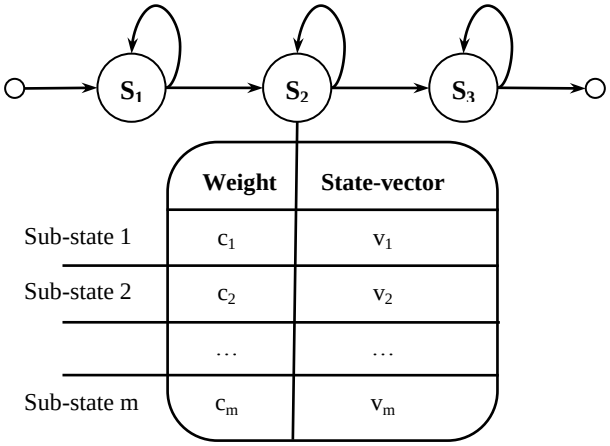

Figure 2. SGMM modeling

Table 6 reports their performances. As expected, SGMM provides better results than the baseline for both types of data.

3.2.3. DNN system

DNN has been known as the big jump in machine learning and is much closed in taking the heir to the ASR throne. For acoustic modeling, DNN replaces the role of GMM. It estimates the posterior for each HMM state. However, in the training phase, DNN still relies on the HMM-GMM structure to determine its target via a force-alignment procedure.

In our Vietnamese implementation, the network is trained using Adam algorithm with a configuration of 0.02 learning rate, 64 mini-batches, and 30 epochs. Its hidden layer count is decided by a tuning phase on the development set as shown in Table 7. For the input layer, speech features (i.e., pitch and MFCC) are composed using a 40 dimensional LDA transformation, and further expanded by concatenating 11 contextual frames. The process ends with a series of 440 dimension vectors as described in [6].

Table 7. DNN tuning

	WER		
# Hidden layers	Broadcast news	Conversations	Overall
4	23.7%	25.4%	24.3%
5	20.5%	23.6%	21.6%
6	19.6%	22.9%	20.8%
7	17.2%	21.8%	18.8%
8	18.1%	22.3%	19.6%

And again, we also explore the effect of fMLLR and discriminative training (with MPE criteria) on DNN. With the numbers outlined in Table 8, DNN surpasses SGMM and the baseline. However, its worth noting that fMLLR gives little improvement for DNN since the network normalizes the speaker effects by its nature. Looking back all the way to the worst overall score of 36.6%, DNN contributes to 66.9% relative improvement, effectively cutting down the error rate to 12.1%.

Table 8. DNN Performances

	WER		
	Broadcast news	Conversations	Overall
DNN	17.2%	21.8%	18.8%
DNN + fMLLR	16.5%	21.2%	18.2%
DNN + fMLLR + MPE	9.7%	16.3%	12.1%

4. CONCLUSIONS

Out of the 4 ASR systems performing on the VN-Corpus, DNN gives best results, obtaining 9.7% WER in the best case. Who could have thought such critical changes in machine learning would bring a strong leap to state-of-the-art Vietnamese ASR. Before the introduction of SGMM and DNN, performances were mediocre. Researchers got stuck in their own limitations, and the works had been stagnant since then.

The outcome of this work implies many possibilities to build sustainable speech applications as well as carry on the research. Viable directions can be bottleneck features and the i-vector approach.

ACKNOWLEDGMENT

This work is part of the research project No.16/2017/HD-KHCNTT, supported by the Institute for Computational Science and Technology, Department of Science and Technology, HCMC-DOST.

REFERENCES

[1] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[2] P. Hoang, *Syllable Dictionary.* Danang Publishing House, 1996.

[3] H. Nguyen *et al.*, "Selection of basic units for Vietnamese large vocabulary continuous speech recognition," in *The 4th IEEE International Conference on Computer Science - Research, Innovation and Vision of the Future,* Ho Chi Minh City, Viet Nam, February 12-16, 2006.

[4] D. Povey *et al.*, "Subspace Gaussian mixture models for speech recognition," in *Proceedings of ICASSP'10*, Dallas, US, 2010.

[5] D. Povey and G. Saon, "Feature and model space feature adaptation with full covariance gaussian," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, US, 2006, pp. 4330–4333.

[6] F. Seide, G. Li, X. Chien, and D. Yu, "Feature engineering in context- dependent deep neural networks for conversational speech transcription," in *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU)*, Hawaii, US, 2011.

[7] Q. Vu *et al.*, "Advances in acoustic modeling for Vietnamese LVCSR," in *International Conference on Asian Language Processing*, Singapore, 2009.

[8] Q. Vu *et al.*, "A robust transcription system for soccer video database," in *International Conference on Audio Language and Image Processing (ICALIP)*, Shanghai, China, 2010.

[9] Q. Vu *et al.*, "isago: The Vietnamese mobile speech assistant for food-court and restaurant location," in *RIVF-VLSP*, Ho Chi Minh City, Viet Nam, 2012.

[10] Q. Vu *et al.*, "A robust Vietnamese voice server for automated directory assistance application," in *RIVF-VLSP*, Ho Chi Minh City, Viet Nam, 2012.

[11] Q. Vu *et al.*, "Temporal confusion network for speech-based soccer event retrieval," in *International Conference on Advanced Technologies for Communications (ATC)*, Ho Chi Minh City, Viet Nam, 2013.

Received on October 05, 2018
Revised on December 04, 2018