A non-expert Kaldi recipe for Vietnamese Speech Recognition System

Hieu-Thi Luong
VNUHCM - University of Science
Ho Chi Minh City, Vietnam
luonghieuthi@gmail.com

Hai-Quan Vu
VNUHCM - University of Science
Ho Chi Minh City, Vietnam
vhquan@fit.hcmus.edu.vn

Abstract

In this paper we describe a non-expert setup for Vietnamese speech recognition system using Kaldi toolkit. We collected a speech corpus over fifteen hours from about fifty Vietnamese native speakers and using it to test the feasibility of our setup. The essential linguistic components for the Automatic Speech Recognition (ASR) system was prepared basing on the written form of the language instead of expertise knowledge on linguistic and phonology as commonly seen in rich resource languages like English. The modeling of tones by integrating them into the phoneme and using the phonetic decision tree is also discussed. Experimental results showed this setup for ASR systems does yield competitive results while have potentials for further improvements.

1 Introduction

Thanks to the improvement by applying deep learning for speech recognition systems (Dahl et al., 2012; Hinton et al., 2012), speech recognition has gained more attention from both research and industrial community. However for minority languages such as Vietnamese, the number of research groups and publications are still limited. One reason for this is the lack of available resources concern of speech recognition and linguistic in general. There are a few attempts to enrich the resources for such languages. One notable example is the GlobalPhone database (Schultz et al., 2013; Schultz and Schlippe, 2014), which provide speech, text data as well as the pronunciation dictionary based on International Phonetic Alphabet (IPA) (International Phonetic Association, 1999) for 20 languages, Vietnamese included. Although for independent researchers it's not feasible to access these public dataset as the cost is quite expensive.

Kaldi (Povey et al., 2011) is an open source Speech Recognition Toolkit and quite popular among the research community. Thanks to the active development, Kaldi is regularly updated with new implementation of state-of-the-art techniques and recipes for speech recognition systems. One motivation for us to define a Vietnamese recipe is to take advantages of such available resources. Another reason is to establish a simple and straightforward Vietnamese recipe so more researchers can start to work on speech recognition system for Vietnamese. In the remain of this paper, Section 2 describes the data we used for the experiments. Section 3 shows the approach for preparing essential linguistic components and describe unique characteristics of Vietnamese languages. Section 4 explains the acoustic modeling and the techniques used to improve the performance of the ASR system. Section 5 evaluates all described recipes while Sections 6 gives a conclusion for our work.

2 Data preparation

2.1 Speech corpus

To evaluate our recipe we prepared a speech corpus by recording speech data from more than 50 native Vietnamese volunteers. For training, 46 speakers (22 males and 24 females) help record 15 hours of speech with 11660 utterances in total. While for testing another set of 19 speakers (12 males and 7 females) recorded 50 minutes of speech with 760 utterances in totals. The recording sessions were conducted in a quiet environment using quality equipments. The size of the corpus is quite small for the current standard of ASR system, however similar size corpus have been commonly used to evaluate Vietnamese ASR system (Quang et al., 2008; Le and Besacier, 2009). We refer to the corpus as VIVOS, more details about the corpus can be found at `http://ailab.hcmus.edu.vn/vivos`.

Table 1: Speech corpus statistics

Set	Speakers	Male	Female	Utterances	Duration	Unique Syllables
Traing	46	22	24	11660	14:55	4617
Testing	19	12	7	760	00:45	1692

2.2 Text corpus

Language model is still an essential part of any state-of-the-art ASR system. In our experiment about 500 MB of text collected from online news and forum from the last 5 years was used to train the decoding language model. The text was first normalized to remove symbols, tags or other non-language elements. Popular abbreviation and numeric expression then replaced by their unrolled written form. The remain text still contained a lot of foreign or incorrect words but was keep as it is and handled in later stage.

3 Linguistic components

3.1 Syllable-based Language Model

In the writing system of Vietnamese language, spaces are used to separate syllables instead of words as in English. Some other languages that share similar trait are Chinese and Japanese. To follow the conventional definition of word in an ASR system one extra step known as word segmentation need to be taken. A Vietnamese word can contain from one up to four syllables, the boundary between words then need to be determined in the segmentation step. Word segmentation is not a trivial task and there is no perfect technique to solve this problem.

For simplification a syllable-based language model can be used instead of word-based. Syllable-based language model helps avoid errors caused by segmentation step and reduce the complexity as the number of syllables is fewer. In our work a trigram syllable-based language model was trained using the text corpus described in previous section. The vocabulary contains 7746 most used Vietnamese-only syllables and the language model was processed to only contain these syllables. As Kaldi would map words and phonemes to their respective integer id, this allows all Vietnamese text to be kept in their Unicode encoding.

3.2 Grapheme-based Pronunciation Dictionary

For Vietnamese ASR systems there isn't a standard dominant pronunciation dictionary. In the GlobalPhone database, IPA was used to construct pronunciation dictionaries for all languages. This unified phoneset open the possibility for multi-language speech recognition system. Although it would be harder to model the unique characteristics of each languages. Another approach is following the Vietnamese phonology definition which suggested that each syllables consist of five components with some components can be redundant. This approach creates a bigger phoneset as it contains diphthong and triphthong. Another approach is using grapheme as phoneme with each character considered as a phoneme as done by Le and Besacier (2009).

In this paper we propose a grapheme-based pronunciation dictionary but not strictly mapping from one character to one phoneme. To simplify the recipe the role and position of each component in syllable are ignored and only two types of phoneme are defined: consonants and vowels. A consonant can be one or up to three characters (instead of one character as grapheme phoneset) while a vowel is a standard vowel with a respective tone. In this setup each tonal variations of a vowel is treated as different phonemes with no relation. To regain tonal information, extra questions could be used to build the phonetic decision tree, the details of this tonal modeling would be discussed in Sections 4.3.

Table 2: Tone integrated grapheme-based phoneset with 99 phonemes in total

Consonants	3 characters	ngh
	2 characters	ch gh gi kh ng nh ph qu tr th
	1 character	b c d đ g h k l m n p r s t v
Vowels	[blank]	a ă â e ê i o ô ơ u ư y
	grave accent	á ắ ấ é ế í ó ố ớ ú ứ ý
	acute accent	à ằ ầ è ề ì ò ồ ờ ù ừ ỳ
	hook	ả ẳ ẩ ẻ ể ỉ ỏ ổ ở ủ ử ỷ
	tilde	ã ẵ ẫ ẽ ễ ĩ õ ỗ ỡ ũ ữ ỹ
	dot below	ạ ặ ậ ẹ ệ ị ọ ộ ợ ụ ự ỵ

It's a well known fact that [ngh] and [ng] is two written form of the same phoneme in Vietnamese. But to keep our setup grapheme-based and easy to follow even for people who do not speak the language, we decided to not combine them and leave such enhancements and other exceptions for future works.

Table 3: Example entries for grapheme-based pronunciation dictionary

nhanh nh a nh	chào ch à o	tôi t ô i
nghiêng ngh i ê ng	ba b a	tối t ố i

4 Systems description

4.1 Acoustic modeling

For acoustic modeling we followed standard recipes of Kaldi. The acoustic features used is 13 dimensions Mel-Frequency Cepstral Coeffiennts (MFCC) with Linear Discriminative Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) applied to 7-splice (3 left and 3 right context) frames and project to a 40-dimensions feature. This feature used to train a conventional triphone GMM acoustic model. Next a discriminative training method Maximum Mutual Information (MMI) was used to train the second model (Povey et al., 2008). A speaker dependent model then trained by applying feature-based Maximum Likelihood Linear Regression (fMLLR) to the acoustic feature (Povey and Saon, 2006). The last model is a hybrid HMM-DNN where DNN was trained to classification the input feature to the corresponding HMM tied states, the fMLLR transformed feature are used to train the hybrid model.

For summary, we trained 4 models: a triphone GMM-based baseline (mGMM), one discriminative trained (mGMM+MMI), one speaker adaptation for GMM (mGMM+SAT) and a hybrid HMM-DNN model with speaker adapted feature (mDNN+SAT). All GMM models contains about 15000 gaussians and 2500 leaves. As for HMM-DNN hybrid setup, DNN contains only 2 hidden layers each with just 300 nodes. The parameters was small when comparing with other Kaldi recipes as the data available for our training is limited.

4.2 Pitch feature for tonal languages

Ghahremani et al. (2014) showed that pitch feature can be helpful for ASR systems especially for tonal languages like Vietnamese and Cantonese. Their implementation for pitch feature extraction is distributed with Kaldi framework. To investigate the effectiveness of pitch in our setup another set of 4 models described above are trained with just one different: pitch feature was augmented into the acoustic features before applied LDA and MLLT.

4.3 Tones clustering using Phonetic Decision Trees

There are some study about modeling tonal information to improve the accuracy of Vietnamses ASR system (Vu and Schultz, 2010; Nguyen et al., 2015) although they are all fall into one of these two approaches: tones are considered as separate phonemes (explicit tone model), or tones are integrated into a phoneme and creating 6 different variation of the same base phoneme (data-driven tone model).

As described in Section 3.2 a tone integrated phoneset is used for our recipes and without further customization our setup is similar to the data-driven tone model (with a slightly different phoneset). To help recreate the relations between the tonal variations of the same base phoneme we utilized the extra questions used to build the phonetic decision tree (Young et al., 1994) to let it asks about tonal questions. This way we can create a more sophisticated modeling for tones and it's also a novel part in our work.

For a typical Kaldi recipe, the question used to build the decision tree are generated automatic based on the tree-clustering of the phones (Povey et al., 2011). Thanks to the flexible structure of Kaldi framework, extra questions about linguistic knowledge can be supplied to further tuning for a particular language. As for Vietnamese we added two simple set of questions about tones: question to group phonemes with same base vowel together and question to group phonemes with the same tone together.

Table 4: Examples of extra questions used to incorporate tones into the decision tree

Same base vowel	Same tone
a á à ả ã ạ ă ắ ằ ẳ ẵ ặ	a ă â e ê i o ô ơ u ư y á ắ ấ é ế í ó ố ớ ú ứ ý

5 Evaluations

Three different recipes followed the description in Section 4 were prepared and evaluated, each with 4 models: the baseline setup using the grapheme-based pronunciation dictionary and standard MFCC feature (baseline), the second recipe with the augmenting of pitch into the acoustic feature (+pitch) and the last one with pitch feature and the incorporation of tones to the phonetic decision tree (+tone).

Table 5: %SyER for 3 recipe each with 4 different models

	baseline	+pitch	+tones
mGMM	19.66	15.14	14.91
mGMM+MMI	18.08	14.96	13.91
mGMM+SAT	15.79	12.07	12.13
mDNN+SAT	13.34	9.54	9.48

Table 5 showed the Syllable Error Rate (%SyER) of three recipes described above. The baseline recipe has a 19.66%SyER for mGMM model and gain 1.58% absolute improvement when training using discriminative method. The mGMM+SAT model trained with the speaker adaptation method fMLLR achieved the best result in all three GMM-based models. While the hybrid HMM-DNN system trained with adapted featured further improve the performance to 13.34 %SyER. This result confirmed the validity of our non-expert recipe for Vietnamese ASR system.

The second recipe with the addition of pitch feature greatly improve the performance in all 4 models. The lowest error is 9.54 %SyER achieved using mDNN+SAT. This once again shows the benefit of using pitch for Vietnamese system. The last recipe with pitch and the incorporation of tones to the decision tree slightly improve the results in the conventional mGMM model and a notable improvement (1% absolute) in mGMM-MMI. However the improvement fades away with 2 models using speaker adaptation technique. This showed the potential of using extra questions for more sophisticated modeling of tones or other linguistic characteristic of Vietnamese.

6 Conclusions

In this work we prepared a toy Vietnamese speech corpus suitable for testing a new speech recognition setup. A grapheme-based Kaldi recipe was established using common information about the language instead of expert knowledge. The tonal information are incorporated into the phonetic decision tree and also show promising result. Similar setups can be constructed for other languages and even if the results do not surpass their counterpart phonetic approach, the using of grapheme-based approach can be useful for tasks like multi-lingual or cross-lingual speech recognition and speech synthesis.

References

George E Dahl, Dong Yu, Li Deng, and Alex Acero. 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42.

Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. 2014. A pitch extraction algorithm tuned for automatic speech recognition. In *Proc. ICASSP*, pages 2494–2498. IEEE.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Viet-Bac Le and Laurent Besacier. 2009. Automatic speech recognition for under-resourced languages: application to Vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, 17:1471–1482.

Thien Chuong Nguyen, Josef Chaloupka, and Jan Nouza. 2015. Study on incorporating tone into speech recognition of Vietnamese. In *Proc. ECMSM*, pages 1–6. IEEE.

Daniel Povey and George Saon. 2006. Feature and model space speaker adaptation with full covariance Gaussians. In *Proc. Interspeech*.

Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon, and Karthik Visweswariah. 2008. Boosted MMI for model and feature-space discriminative training. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4057–4060. IEEE.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. In *Proc. IEEE ASRU*. IEEE Signal Processing Society.

Nguyen Hong Quang, Pascal Nocera, Eric Castelli, and Trinh Van Loan. 2008. A novel approach in continuous speech recognition for Vietnamese, an isolating tonal language. *Proc. SLTU*.

Tanja Schultz and Tim Schlippe. 2014. Globalphone: Pronunciation dictionaries in 20 languages. In *Proc. LREC*, pages 337–341.

Tanja Schultz, Ngoc Thang Vu, and Tim Schlippe. 2013. Globalphone: A multilingual text & speech database in 20 languages. In *Proc. ICASSP*, pages 8126–8130. IEEE.

Ngoc Thang Vu and Tanja Schultz. 2010. Optimization on Vietnamese large vocabulary speech recognition. In *Proc. SLTU*, pages 104–110.

Steve J Young, Julian J Odell, and Philip C Woodland. 1994. Tree-based state tying for high accuracy acoustic modelling. In *Proc. ARPA Human Language Technology Workshop*, pages 307–312. Association for Computational Linguistics.