The Vietnamese Speech Recognition Based on Rectified Linear Units Deep Neural Network and Spoken Term Detection System Combination

Shifu Xiong, Wu Guo, Diyuan Liu

National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China

domine@mail.ustc.edu.cn, guowu@ustc.edu.cn, ldy2012@mail.ustc.edu.cn

Abstract

In this paper, we report our recent progress on the under-resource language automatic speech recognition (ASR) and the following spoken term detection (STD). The experiments are carried on the National Institute of Standards and Technology (NIST) Open Keyword Search 2013 (OpenKWS13) evaluation Vietnamese corpus. Compared with the conventional ASR system, we made the following modifications to improve recognition accuracy. First, pitch features and tone modeling are applied to cover pitch and tone information since Vietnamese is a tonal language. Second, automatic question generation for decision tree is used for state tying to address the problem of lack of linguistic knowledge. Finally, we investigate rectified linear units (ReLUs) activation function and cross-lingual pre-training in deep neural network (DNN) acoustic model training. In the STD procedure, we adopt term-dependent score normalization and combine the outputs of diverse ASR systems to increase actual term weighted value (ATWV). After applying these methods, our current best single system achieves 48.32% word accuracy and 0.398 ATWV after STD system combination on OpenKWS13 Vietnamese development set.

Index Terms: under-resource speech recognition, deep neural network, rectified linear units, spoken term detection, system combination

1. Introduction

STD is a technology that locates a specified, potentially multi-word keyword from archives of speech data quickly and precisely [1]. In 2006, NIST held the first international STD evaluation. In this evaluation, researchers achieved a very high performance on large-resource language (e.g. English, Mandarin) and clean speech conditions [1][2]. However, it's still a challenge problem on noisy speech for under-resource language. In 2013, NIST organized a similar evaluation program called Open Keyword Search Evaluation (OpenKWS) again [3]. It's an extension of the 2006 Spoken term detection evaluation. But the goal of the program is to reduce the difficulty of building high-performing STD systems on a new language quickly with significantly less training data that are also much noisier and more heterogeneous than what has been used in the current state-of-the-art. The surprise language is Vietnamese in OpenKWS13.

The general framework of a STD system is made up of two subsystems: an automatic speech recognition (ASR) subsystem and a term detection subsystem, the former converts speech signals to word or sub-word lattices which are subsequently indexed, and the latter searches keywords in the index and ascertains the detection candidates. A good ASR system is essential for building STD system but not sufficient. Recently, it has been demonstrated that significant improvement on STD task can be obtained by diverse of STD systems combination [4]. In this paper, as a participant of OpenKWS13, we focus on how to build a powerful Vietnamese ASR and STD system. The main contribution of the paper is that we present our experience and some universal technologies for under-resource speech recognition and describe some important methods for STD. Vietnamese is a tonal language, so pitch features are appended to the conventional PLP features and the tonal phonemes are selected as the acoustic modeling units. A similar work is presented in [5]. Decision tree based state tying is a baseline feature of state-of-art speech recognition systems. However, for a new language, we have no linguistic expert to define proper phonetic questions for state tying. To address this problem, we investigate automatic generation of questions [6] and compare it with data-driven based state tying. We also present that building multi noise models is a good solution to fit a variety of noise rather than using a general model. At last, cross-lingual deep neural network (DNN) pre-training shows its dramatic power in closing the performance gap between large-resource and low-resource language [7-9]. It helps neural network to learn more information than mono-lingual training and prevents network from overfitting. Additionally, rectified linear units (ReLUs) have already been successfully introduced in speech recognition and achieved modest improvement compared with sigmoid by using dropout to avoid overfitting [10]. In this paper, we combine ReLUs with cross-lingual training to improve recognition performance rather than dropout. In the STD procedure, we investigate term-dependent confidence normalization and system combination. Ideal STD system should have accurate scores and high recall for every possible keyword. Although the results of [11] indicate that term-dependent confidence measure provides only weak improvement for in-vocabulary terms, we still believe that it can work well for STD with low word accuracy speech recognizer. ATWV, the metric used for STD task, emphasizes recall of rare terms. When speech recognizer's performance is poor, it often happens that some keywords have only a few detection candidates and their scores are very low. Such detection candidates are normally rejected by decision maker. Term-dependent normalization is such a technique to solve this issue. Furthermore, considering the complementarities among different ASR systems, STD system combination can also help to increase ATWV [4,13,14].

The remainder of this paper is organized as follow. In Section 2, we give a detail description of the optimization methods for Vietnamese ASR, including extraction of pitch information, automatic generation questions, cross-lingual training and rectified linear units, and some our experiences are also presented in this section. In Section 3, term-dependent score normalization

978-1-4799-4219-0/14/$31.00

and system combination techniques are described. Experiments and results are presented in Section 4. Finally, Section 5 concludes the paper.

2. Key Techniques for Speech Recognition

As an important part of STD, the performance of ASR has significant influence on STD. In this section, we present some key techniques to improve performance of Vietnamese speech recognition, including optimization methods for front ends, acoustic models and DNN training.

2.1. Pitch Features and Tone Modeling

In so-called tonal languages, e.g. Mandarin, pitch feature and tone modeling are a critical component for large vocabulary continuous speech recognition systems. Vietnamese is also a tonal language, so we adopt a similar configuration for Vietnamese speech recognition as Mandarin ASR systems. The pitch features used in this paper include fundamental frequency (F0) and its first and second order derivatives as well as the degree of voicing. In this work, we extract pitch features using the approach described in [12]. We compute the candidate F0 of each frame using subharmonic summation, and then we compute the voicing degree of each frame after obtaining F0. Furthermore, we use dynamic program to eliminate discontinuities in the pitch track caused by multiple and sub-multiple gross pitch errors and smooth F0 of silence and unvoiced frames. Finally, 4-dimensional pitch features are appended to the original 39-dimensional PLP feature vector (13th order PLPs and their first and second derivatives). Additionally, tones are used to represent phone level distinctions. In the tone modeling, phonemes with different tone are treated as different acoustic units. Each vowel of Vietnamese has six tones. For example, the vowel 'a' is represented by the models 'a1', 'a2', , 'a6'.

2.2. Automatic Generation Question Set

As a surprise language, we can't design a suitable Vietnamese question set for state tying since we lack linguistic knowledge. In this condition, data-driven is a natural choice to tie states, but data-driven based state tying can't deal with unseen triphones. Therefore, we need an algorithm of automatic clustering and generation of contextual questions for tied states. In this work, we generate question set using the approach described in [6]. It's a clustering technique based on likelihood maximization criteria in which the first and second half states of context independent models are used to generate right-context and left-context questions. In the clustering procedure, groups of phones are recursively clustered until only two maximally-separated clusters and it's repeatedly performed on each of these clusters, followed by exhaustive partitioning. The procedure stops unless the number of phone in both the two maximally-separated clusters is less than or equal to 2. All the maximally-separated clusters pairwise generated in the clustering procedure are a part of final question set.

2.3. Cross-lingual Training and Rectified Linear Units

For speech recognition, the network structure is very deep and a lot of training data is required for parameters estimation. A frequently-used network in ASR, for example, has six or more hidden layers and each hidden layer has 1k$\sim$2k units. It has enormous parameters and can easily overfit for under-resource speech recognition condition. Additionally, neural network updates parameters with stochastic gradient descent based back propagation algorithm and may fall into a local optima when data is not enough. Starting from an existing large-resource language network, cross-lingual training helps to prevent overfitting and eavesdrop knowledge from large-resource language information. In this paper, the large-resource language is Mandarin and has 1000h training data, and the Mandarin DNN parameters are used as Vietnamese pre-training DNN. Since we have about 80 hours Vietnamese training corpus, all the neural network parameters rather than only the softmax layer are updated in fine-tuning procedure.

In additional, we replace sigmoid activation function of DNN with ReLU and train it in the cross-lingual training schedules. This approach is inspired from [10]. In [10], with dropout for avoiding overfitting, ReLUs provide a modest improvement over sigmoid. Cross-lingual training can also prevent overfitting. Therefore, we combine ReLUs activation function in our training.

2.4. Experiences for Building ASR System

Here, we share our experiences for building ASR system.

1. Tone modeling results in a much larger number of triphone lists if we extend unseen triphones by traversing all the phoneme combinations. It results in a serious burden for both training acoustic modeling and decoding. Adding unseen triphones extended by dictionary can alleviate this problem.
2. For the noisy modeling, using multi noise models is a good solution to fit a variety of noise speech rather than using a general model.
3. The more data the large-resource language has, the better performance the cross-lingual training will be obtained. Even if the data of the source language is same as the target language, cross-lingual training is also better than initialization from pre-training.

3. Score Normalization and System Combination

In this section, we investigate term-dependent score normalization and system combination that contribute to increase ATWV. According to ATWV definition, the benefit of correctly finding a term is inversely proportional to the term frequency while the cost of a false alarm is almost independent to the term frequency; therefore, it emphasizes recall of rare terms.

Sum-to-one normalization method works well in [13][14], but it's still a term-independent confidence that does not consider the evaluation metric ATWV. Term-dependent score normalization is an ATWV-oriented normalization method. It requires the expected benefit of a putative detection is positive; otherwise the decision maker will reject it. Simply given a term w with scores $s_{w,1}, s_{w,2}, \ldots, s_{w,n}$, the normalized scores become

$$s'_{w,i} = \frac{s_{w,i} \times \alpha + \gamma}{\sum_{j=1}^{n} s_{w,j}} - \beta \frac{1 - s_{w,i} \times \alpha - \gamma}{T - \sum_{j=1}^{n} s_{w,j}} \qquad (1)$$

where α, γ are two adaptable parameters to compensate for any bias, T is the total number of trials (e.g. seconds in the audio), β is a constant, both T and β come from the definition of ATWV. The sum of all the posterior scores represents an approximation of the number of occurrences of the term. Note that the formula boosts the scores of terms with generally low scores; therefore, for rare terms, the probability of missed detection will be lower.

All fusion methodologies are motivated by the desire to emphasize the reliability of detection candidates appearing in multiple systems. For system combination, we use the approach described in [14], named MTWV-weighted CombMNZ (WCombMNZ). Meta-hit is formed by merging hit lists of all single systems. Denote h_i the hit that contributes to the meta-hit H from the i-th system among n systems, $s(h_i)$ the score of the hit h_i, m_H is the number of indices having a non-zero score for this meta-hit H and $MTWV_i$ is the maximizing ATWV of i-th system. WCombMNZ computes score of the meta-hit H as

$$m_H \times \sum_{i}^{n} \frac{MTWV_i}{\sum_{j=1}^{n} MTWV_j} \cdot s(h_i), \quad (2)$$

In this paper, term-dependent score normalization is applied after system fusion.

4. Experiments

In this paper, our experiments are conducted on the OpenKWS13 Vietnamese Full Language Pack (FullLP).

4.1. Data and Description

The training set used for building ASR system contains 80 hours speech. The data covers a broad speaker population, and includes a variety of dialects and speaker ages, and is approximately gender-balanced. In addition, a significant portion of the audio data is labeled as non-lexical speech events or non-speech events, e.g., hesitations, breath, laugh. The 10K dictionary is used as the pronunciation dictionary for the system. The total number of phonemes is 70. There are 25 consonants and 45 vowels (12 monophthongs, 25 diphthongs and 8 triphthongs), and each vowel has six tones. The development set (dev-set) is 10-hours speech. We use 2-hours of randomly chosen speech from dev-set for testing speech recognizers word accuracy. Keyword search is performed on the all 10-hours dev-set, with an official keyword list provided by NIST for the STD evaluations. The keyword set contains 4065 queries.

4.2. Speech Recognition

Speech recognition experiments are performed using HTK. The 2-hours test data is decoded using HDecode decoder, and we made some modifications to fit for DNN decoding. A bigram language model, built using the SRILM tools with the transcripts of training set, is used in decoding. The HResult tool is used to evaluate the recognition performance.

4.2.1. Traditional GMM-HMMs System

First of all, we build the GMM-HMMs baseline system using the regular 39-dimension PLP features, data-driven based state tying, a general model for all the non-lexical or non-speech events. The baseline models are trained based on maximum likelihood estimation (MLE), including 5005 tied states and 25 Gaussian components per state. The baseline system's performance is listed in the row 2 of Table 1, and it is very poor. Then we improve the baseline performance using the methods described in section 2, including pitch extraction, tone modeling, automatic generation question set and multi noise modeling. As we can see from the third row of table 1, compared with baseline system, the 43-dimensional features (PLP + pitch) give a 2.69% absolute improvement; therefore the 43-dimensional features are adopted in the following systems. Furthermore, the tone modeling gives a 4.11% absolute improvement against the performance of the row third. With multi noise model and automatic question set generation, the GMM-HMMs system can achieve word accuracy about 29.13%.

Table 1: Recognition performance of different optimization methods for traditional GMM-HMM.

Method	word accuracy (%)
baseline	17.01
+ pitch feature	19.70
+ tone modeling	23.81
+ automatic question set generation	27.71
+ multi noise models	29.13

4.2.2. Deep Neural Network Based System

Two different DNNs are trained separately. The first DNN is trained on a 9-frame context window of 43-dimension PLP features, with 6 hidden layers and a tied state output layer. This results in an architecture of 473-2048-2048-2048-2048-2048-2048-5005. We call DNN hybrid model based on this kind of architecture DNN-HMMs. The second DNN has 4 wide hidden layers and a bottleneck layer, resulting in an architecture of 473-2048-2048-2048-2048-43-5005. This DNN is used for bottleneck features extraction. We use bottleneck features to train another GMM-HMMs described in the above section. We name it BN-GMM-HMMs. Mono-lingual training is done with pre-training initialization. Cross-lingual training is done with initialization from 1000-hours Mandarin network. All the DNNs are trained with the standard back propagation algorithm using a cross entropy error criterion. The learning rate and stopping criterion are controlled by the frame classification error on a cross validation data set.

Tabel 2 shows the recognition accuracy using mono-lingual training and cross-lingual training based on sigmoid and ReLUs. The performance of the baseline system of table 2 is extracted from Table 1. The mono-NN means neural network trained with mono-lingual training, cross-NN means neural network trained with cross-lingual training, and ReFA means alignments are regenerated from the system of the fifth row using DNN models to retrain DNNs, the second column in Table 2. The performance in the brackets is optimized using lattice-based sequential feature space discriminative training method to extract more discriminative bottleneck features follow by model space discriminative training based on minimum phone error criterion [15].

As we can see from Table 2, ReLU is superior to sigmoid. With cross-lingual training, the performance gap between them is larger. The best performance can be obtained by combining cross-lingual , ReFA and ReLUs. The decoding lattice of the ReLUs based DNN-HMMs and BN-GMM-HMMs systems will be used for subsequent STD.

4.3. Spoken Term Detection Results

After decoding, we convert the dev-set to lattices and then the consensus networks (CNs) are generated form the lattices. Keyword search is performed in the CNs index. The baseline STD system is our submitted system to OpenKWS13. It is a single system and the recognizer's performance is very poor (word recognition accuracy 29.70%). BN and DNN are the best sys-

Table 2: Word recognition accuracy (%) with different network configurations.

	DNN-HMMs		BN-GMM-HMMs	
	Sigmoid	ReLUs	Sigmoid	ReLUs
Baseline	29.13			
Mono-NN	41.40	42.86	–	–
Cross-NN	44.30	46.29	41.98	–
+ ReFA	45.60	46.93	43.57	44.25 (48.32)

tem presented in Table 2. "Sys Comb" means BN and DNN system fusion using method in section 3. STO means sum-to-one and TD means term-dependent. Experiments in Table 3 show that high recognition performance and score normalization and system combination can significantly increase ATWV, term-dependent score normalization work well in low word recognition condition and is slightly better than sum-to-one normalization.

Table 3: ATWV results with score normalization and system combination on OpenKWS13 dev-set.

Norm Method	Baseline	BN	DNN	Sys Comb
–	0.102	0.294	0.287	–
STO	0.132	0.347	0.337	0.391
TD	0.155	0.355	0.345	0.398

5. Conclusions

In this paper, we show that cross-lingual DNN training with ReLUs activation function can achieve an absolute 19.2% recognition performance over ML-trained GMM-HMMs. Our best BN-DNN-HMMs system can obtain a word accuracy of 48.32%. The term-dependent score normalization followed by system combination approach improves by relative 12.1% STD performance, achieving an ATWV of 0.398. In addition, we present some key techniques for ASR. Among them, automatic generation of question set and cross-lingual DNN training are language independent and are particularly well suited for under-resource speech recognition.

6. References

[1] J. Fiscus, J. Ajot, J. Garofolo, and G. Doddington, "Results of the 2006 Spoken Term Detection Evaluation," in *Proc. 2007 SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 51-57.

[2] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in *Proc. Interspeech*, 2007, pp. 2393-2396.

[3] "Openkws13 keyword search evaluation plan," http://www.nist.gov/itl/iad/mig/upload/OpenKWS13-EvalPlan.pdf

[4] L. Mangu, H. Soltau, H. K. Kuo, B. Kingsbury, and G. Saon, "Exploiting diversity for spoken term detection," in *Proc. ICASSP*, 2013, pp. 8282-8286.

[5] N. T. Vu, and T. Schultz, "Vietnamese large vocabulary continuous speech recognition," in *Automatic Speech Recognition and Understanding, 2009. ASRU 2009. IEEE Workshop on. IEEE*, 2009, pp. 333-338.

[6] R. Singh, B. Raj, and R. M. Stern, "Automatic clustering and generation of contextual questions for tied states in hidden Markov models," in *Proc. ICASSP'99*, 1999, pp. 117-120.

[7] J. T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013, pp. 7304-7308.

[8] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. "Multilingual acoustic models using distributed deep neural networks," in *Proc. ICASSP, 2013*, pp. 8619-8623.

[9] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proc. ICASSP, 2013*, pp. 7319-7323.

[10] G. E. Dahl, T. N. Sainath and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. ICASSP*, 2013, pp. 8609-8613.

[11] D. Wang, S. King, J. Frankel, and P. Bell, "Term-dependent confidence for out-of-vocabulary term detection," in *Proc. Interspeech*, 2009, pp. 2139-2142.

[12] C. H. Huang and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," in *Proc. ICASSP*, 2000, pp. 1523-1526.

[13] D. Karakos, R. Schwartz, S. Tsakalidis, L. Zhang, S. Ranjan, T. Ng, et al. "Score normalization and system combination for improved keyword spotting," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE*, 2013, pp. 210-215.

[14] J. Mamou, J. Cui, X. Cui, M. J. F. Gales, B. Kingsbury, K. Knill, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schluter, A. Sethy, and P. C. Woodland, "System combination and score normalization for spoken term detection," in *Proc. ICASSP*, 2013, pp. 8272-8276.

[15] D. Y. Liu, S. Wei, W. Guo, Y. B. Bao, S. F. Xiong, and L. R. Dai, "Lattice based optimization of bottleneck feature extractor with linear transformation," in *Proc. ICASSP*, 2014. To appear.