The NNI Vietnamese Speech Recognition System for MediaEval 2016

Lei Wang[1], Chongjia Ni[1], Cheung-Chi Leung[1], Changhuai You[1], Lei Xie[2], Haihua Xu[3], Xiong Xiao[3], Tin Lay Nwe[1], Eng Siong Chng[3], Bin Ma[1], Haizhou Li[1,4]
[1]Institute for Infocomm Research (I²R), A*STAR, Singapore
[2]Northwestern Polytechnical University (NWPU), Xi'an, China
[3]Nanyang Technological University (NTU), Singapore
[4]National University of Singapore (NUS), Singapore
{wangl,ccleung}@i2r.a-star.edu.sg, lxie@nwpu.edu.cn, {haihuaxu,xiaoxiong}@ntu.edu.sg

ABSTRACT
This paper provides an overall description of the Vietnamese speech recognition system developed by the joint team for MediaEval 2016. The submitted system consisted of 3 sub-systems, and adopted different deep neural network-based techniques such as fMLLR transformed bottleneck features, sequence training, etc. Besides the acoustic modeling techniques, speech data augmentation was also examined to develop a more robust acoustic model. The I²R team collected a number of text resources from the Internet and made them available to other participants in the task. The web text crawled from the Internet was used to train a 5-gram language model. The submitted system obtained the token error rate (TER) of 15.1, 23.0 and 50.5 on *Devel local* set, *Devel* set and *Test* set, respectively.

1. INTRODUCTION
The zero-cost speech recognition task [1] at MediaEval 2016 aims to build Vietnamese automatic speech recognition (ASR) systems using publicly available multimedia resources (such as texts, audios, videos, and dictionaries). About 10 hour transcribed speech data was provided by the task organizers. The provided data came from 3 different sources and were recorded in different environments.

Our submitted system consists of 3 sub-systems: 1) a DNN-HMM system with MFCC; 2) a DNN-HMM with data augmentation; 3) a DNN-HMM with bottleneck features (BNFs). To build the acoustic models, tonal information was involved in the front-end processing, and bottleneck features (BNFs) were used. Traditional GMM-HMM models were used to align the speech frames to the phonetic transcription, and Deep Neural Network (DNN) models were trained using cross-entropy criterion, followed by sequence training based on state-level minimum Bayes risk (sMBR). To improve the robustness of our acoustic model, data augmentation was attempted. To build a language model (LM), web page data were crawled from the Internet. Other publicly available text resources were also involved, and we made them to be accessible by all the participants.

2. DATA CONTRIBUTION
The I²R team collected and contributed a number of text resources listed as below:

- 890 thousand URLs of Vietnamese web pages crawled using a large number and variety of keywords and phrases;
- An XML dump of Vietnamese Wikipedia's articles [2] and its cleaned text;
- 4 Vietnamese word lists [3] of different sizes.

The above text data were made available to other participants in the task. The corresponding data pack, i2r-data-pack, is also available for download at https://github.com/viet-asr/i2r-data-pack-v1.0.

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

3. APPROACHES
This section describes the acoustic modeling of the 3 sub-systems, as well as the text data and lexicon used for language modeling. The 3 sub-systems share the same lexicon and language model in decoding, which are described in Section 3.4. The hypotheses of the 3 sub-systems were fused to a single system for the final submission using the ROVER algorithm [4].

3.1 DNN-HMM System with MFCC
We used 56-dimensional acoustic features consisting of 13-dimensional MFCC, 1-dimensional F0, and their derived deltas, acceleration and third-order deltas, as the input of a DNN-HMM hybrid system [5]. The acoustic model considers 94 graphemes which were discovered from the training transcription as monophones. The context-dependent triphones were modeled by 801 senones. The final model was trained with cross-entropy criterion and sMBR [6], on top of a GHM-HMM model trained using maximum mutual information (MMI) [7]. The DNN structure consists of 5 layers with 1024 nodes per layer. The total duration of the training corpus is about 10 hours, and it is provided by the organizer.

3.2 DNN-HMM System with Data Augmentation
Among the 10 hours of Vietnamese training data, some utterances are relatively clean, some have been filtered through denoising algorithms, and most of them are contaminated by different kinds of background noise. To improve the robustness of our recognition system against noisy speech, we augmented the training utterances by corrupting each original utterance with noise and applying speech enhancement on each original utterance. After data augmentation, the total amount of training data was increased by 2 times.

Different kinds of background noise were extracted from the training utterances using a voice-activity-detection algorithm. Representative noise segments were selected and randomly added into the original training utterances. Speech enhancement includes

two main estimation modules: the estimation of speech and the estimation of noise. We used the modified version of log-spectral-amplitude (LSA) minimum mean square error (MMSE) algorithm as the speech estimator [8]. The quality of estimated speech with the same speech estimator heavily depends on the accuracy of the estimation of the noise statistics. To improve the performance in non-stationary background noise condition, we adopted a minimum searching with the speech presence probability (SPP) for noise estimation [9].

With the augmented training data, we trained another DNN-HMM hybrid system with the same network structure (i.e. the same numbers of hidden layers, hidden units per layers, and tied states) as the system in Section 3.1. During recognition, we used the original development/test utterances for decoding.

3.3 DNN-HMM System with BNFs

Another DNN-HMM hybrid system used bottleneck features (BNFs) and fMLLR features as its input. This type of BNF-based systems [10-11] is commonly used in the limited training data condition. In the bottleneck feature extraction, 13-dimensional MFCC and 2-dimensional F0-related features were extracted. Nine adjacent frames of features were then concatenated and applied with LDA+MLLT+fMLLR transform. MLLT makes the features to be better modeled by diagonal-covariance Gaussians. The resultant 40-dimensional fMLLR features were used for BNF extraction from a DNN consisting of 6 hidden layers with 1024 nodes in each non-bottleneck layer, and 42-dimensional BNFs were obtained.

The 42-dimensional BNFs and the 40-dimensional fMLLR features were then concatenated to form 82-dimensional features. Then fMLLR transform was applied again (60-dim) to normalize inter-speaker variability. The final 60-dimensional features were used as the input of another DNN. This DNN contains 6 layers, each layer contains 1024 nodes, and the output layer contains 2073 senones. The final model was trained with cross entropy criterion and sMBR, on top of a GHM-HMM model trained using MMI.

3.4 Lexicon and Language Model

The grapheme-based lexicon contains about 11,000 Vietnamese syllables and English words which occur in the training transcription and the 74,000 Vietnamese word list in i2r-data-pack.

A 5-gram LM was trained using the following 4 data sources:

1) 7GB of text extracted from the list of web pages in i2r-data-pack;
2) 750MB of text from Wikipedia's articles in i2r-data-pack;
3) 90MB download of Vietnamese-English subtitles released by BUT;
4) Transcription of training utterances.

The final LM was obtained by linear interpolation of four LMs, each of which was trained using one of the above data source. The interpolation weights were optimized using the transcript of the development data set (*Devel local*). Perplexity and TER were reduced on *Devel local* set in our preliminary systems when the web data were included in the language model training.

4. RESULTS AND DISCUSSION

Table 1: ASR performance of different sub-systems and the fused system

	LM	TER (%)		
		Devel local	*Devel*	*Test*
MFCC-DNN-HMM	11K lexicon 5-gram LM	17.4	26.9	55.1
+ Data Augmentation		18.4	26.5	53.9
BNF-DNN-HMM		18.5	25.5	50.9
Fusion of 3 systems		**15.1**	**23.0**	**50.5**

Table 1 summarizes the ASR performance of each system on 3 different test sets. We observed that the performance on *Devel local* set is not inconsistent to that on the other 2 sets. We believe that it is because of the small amount of data (~13 minutes) involved in *Devel local* set. Moreover, since *Devel local* set is a small sub-set of *Devel* set, our analysis will focus on *Devel* and *Test*. However, note that most of system configurations were tuned on the *Devel local* set, to avoid the frequent upload of our results to the leader board.

The BNF-based system obviously has the best performance among all sub-systems probably due to the contribution of more robust bottleneck features and speaker normalization by fMLLR. The data augmentation technique improves the system performance by relatively 1.5% and 2.2% on *Devel* and *Test* sets, respectively. Data augmentation provides varieties of training speech data with noisy background so that it improves the robustness of the acoustic model. Moreover, we believe that the resultant acoustic model is more robust against unseen data, e.g. the surprise data in the *Test* set.

The fused system has the overall best performance, and it can be attributed to that the 3 sub-systems are complementary.

5. CONCLUSION

This work describes the acoustic modeling of 3 sub-systems and the approach for language modeling under the limited training data condition. We relied on the provided training corpus to build acoustic models, and effort was made to collect web text data to build an LM.

We reported the ASR performance which was achieved by the deadline of the task. In the future work, we will examine the data augmentation on the BNF-based system. We will further investigate to use the speech data contributed by other participants.

REFERENCES

[1] Igor Szoke and Xavier Anguera, "Zero-Cost Speech Recognition Task at Mediaeval 2016," in *Proc. MediaEval 2016 Workshop*, Hilversum, Netherlands, Oct. 2016.

[2] https://dumps.wikimedia.org/viwiki/20160501/viwiki-20160501-pages-meta-current.xml.bz2

[3] http://www.informatik.uni-leipzig.de/~duc/software/misc/wordlist.html

[4] J. Fiscus, "A post processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. ASRU 1997*, 1997, pp. 347- 354.

[5] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol.20, no.1, pp. 30-42, Jan. 2012.

[6] Veselý, Karel, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Interspeech 2013*, pp. 2345-2349, 2013.

[7] D. Povey, "Discriminative training for large vocabulary speech recognition," *Ph.D. dissertation*, Cambridge University Engineering Dept, 2003.

[8] R. Gemello, F. Mana, and R. D. Mori, "Automatic Speech Recognition with a Modified Ephraim-Malah Rule," in *IEEE Signal Processomg Letters*, vol.13, no.1, pp. 56-59, Jan. 2006.

[9] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol.48, pp. 220-231, 2006.

[10] Chongjia Ni, Cheung-Chi Leung, Lei Wang, Nancy F. Chen, and Bin Ma, "Unsupervised data selection and word-morph mixed language model for Tamil low-resource keyword search," in *Proc. ICASSP 2015*, Brisbane, Australia, April 2015.

[11] Chongjia Ni, Cheung-Chi Leung, Lei Wang, Haibo Liu, Feng Rao, Li Lu, Nancy F. Chen, Bin Ma, and Haizhou Li, "Cross-lingual deep neural network based submodular unbiased data selection for low-resource keyword search," in *Proc. ICASSP 2016*, Shanghai, China, March 2016.