BUT Zero-Cost Speech Recognition 2016 System Description

Miroslav Skácel, Martin Karafiát, Lucas Ondel, Albert Uchytil, Igor Szöke
BUT Speech@FIT
Brno University of Technology, Czech Republic
{iskacel, karafiat, iondel, xuchyt03, szoke}@fit.vutbr.cz

ABSTRACT

This paper describes our work on developing speech recognizers for Vietnamese. It focuses on procedures to prepare provided data precisely. We aim on analysis of the textual transcriptions in particular. Methods to filter out defective data to improve performance of final system are proposed and described in detail. We also propose cleaning of other textual data used for language modeling. Several architectures are investigated to reach both sub-tasks goals. The achieved results are discussed.

1. INTRODUCTION

For the Zero-Cost 2016 Speech Recognition task, we developed one Large Vocabulary Continuous Speech Recognition (LVCSR) system and one subword system for on-time submission and two more LVCSR systems for late submission. LVCSR systems were based on our previous knowledge from Babel Program [1][2]. We presented two types of LVCSRs. The first type uses Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Deep Neural Network (DNN) from Babel 2014 [1]. The second one adopted Bidirectional Long Short-Term Memory (BLSTM) approach from Babel 2016 [2]. Our goal was to modify and apply existing Babel LVCSR systems for this year's target language that was Vietnamese. For subword sub-task, we exploited acoustic unit discovery model. See [3] for more details on each of the sub-tasks.

2. DATA PREPARATION

The given mix of audio data, transcripts and additional texts were preprocessed and elementary cleaned before training of our systems.

2.1 Audio

For BLSTM system, the original 16kHz audio was used. For GMM/DNN based system, the original audio was downsampled from 16kHz to 8kHz to fit our training scripts. We also used the information about the audio length to process transcription texts later.

2.2 Transcriptions

All other symbols but letters from Vietnamese alphabet (punctuation marks, brackets, etc.) were removed from transcriptions and text was converted to uppercase.

Numerals composed of digits were expanded to its textual form. We took textual transcription of basic numerals (0,1,2,..,100,1000,...) in Vietnamese language. The procedure to compose a number in Vietnamese is simple compared to some other languages and follows very logical rules. Thus, the textual translation was iteratively created for every number composed of digits.

A transcription was a simple text file for each of audio files. There was no information about transcription alignment so it could match the whole audio. Audio files longer than 1 minute were discarded from first iteration of LVCSR training (described in 3.1) due to high memory demands. We used the alignment obtained during training to split transcriptions into smaller segments. For every detected silence longer than 0.5 s, the segment was divided. If the segment lasted longer than 15 seconds, it was also split by first possible silence detected regardless the duration. This allowed us to utilize the whole training data set with acceptable memory demands during training.

In the next step, we focused on defective audio and improper transcription texts. The average log-likelihood of speech frames was calculated by accumulating the log-likelihood from the first iteration system for all speech frames and dividing by the number of frames in the given audio. The same was done for silence frames. The log-likelihood of speech was very low when the audio contained a silence/noise only, the transcription did not strongly correspond to the audio or a part of the transcription was missing. Therefore, we discarded defective files from further training by ad-hoc threshold set to -100. We used 92 % of training data after cleaning.

2.3 Language Models (LMs)

Three LMs were created for this sub-task. The first LM for on-time submitted LVCSR system was trained on text taken from training set transcriptions.

The second LM was trained on Vietnamese subtitles. We took provided wordlist to create set of Vietnamese letters to filter out words from other languages. Again, punctuation and quotation marks, brackets and other symbols were eliminated from the text. Numerals composed of digits were transformed to textual notation in the same way we did previously. Sentences comprising less than 3 words were discarded as well. The text was converted to uppercase.

We were provided with a set of URLs which headed to the websites in Vietnamese. We extracted the inner text from all of the HTML tags. However, the data contained a lot of unusable text. We removed the lines containing any special chars and numbers at first. After that, we created a wordlist from already cleaned up data and did a filtering according to

MediaEval 2016 Workshop, October 20-21, 2016, Hilversum, Netherlands

System	Devel all (ELSA / Forvo / RhinoSpike)	Test all (ELSA / Forvo / RhinoSpike / YouTube)
P-BUT - Babel Kaldi BLSTM 16kHz	17.9 (6.4 / 58.1 / 15.8)	48.0 (4.9 / 55.7 / 35.4 / 87.2)
L-BUT - Babel Kaldi BLSTM 16kHz - LM tune	17.6 (6.2 / 56.4 / 16.9)	46.3 (4.6 / 52.6 / 32.2 / 84.7)
L-BUT - Babel GMM/DNN 8kHz	36.1 (29.7 / 68.5 / 23.4)	55.7 (28.0 / 59.3 / 44.9 / 81.4)

Table 1: *Results of the LVCSR systems for overall score and single test subsets (shown in parentheses) in WER metric. System labeled by* P *was submitted on-time;* L *denotes late submission systems.*

System	Devel all (ELSA / Forvo / RhinoSpike)	Test all (ELSA / Forvo / RhinoSpike / YouTube)
P-BUT AUD phone-loop	5.08 (6.45 / 8.76 / 14.19)	4.56 (5.52 / 9.59 / 18.49 / 7.59)

Table 2: *Results of the subword system for overall score and single test subsets (shown in parentheses) in NMI metric.*

it. The duplicate lines were removed. In total, we obtained about 460k sentences to create our third LM.

These three LMs were combined together in a linear way (denoted as *LM tune* in Table 1).

3. LVCSR SYSTEMS

We developed two different LVCSR system for the first sub-task - GMM/DNN and BLSTM architectures.

3.1 GMM/DNN

The automatic speech recognition (ASR) system developed for Babel [1] focuses on languages with limited amount of training data. This architecture uses Stacked Bottle-Neck Neural Network (SBN NN) for feature extraction that overcomes standard Bottle-Neck features. It contains two consecutive NNs. The first one has four hidden layers with 1500 units each except the bottle-neck layer. The BN layer is the third hidden with 80 neurons. It outputs 21 frames that are downsampled and taken as an input to the second NN. This NN has the same structure. The bottle-neck layer consist of 30 neurons. It outputs SBN features that are used to train GMM-HMM system.

This HMM-based speech recognition system works with tied-state triphones and uses standard maximum likelihood technique for training. Word transcriptions are get using 3-gram LM taken from cleaned training texts.

To perform speaker adaptation, we trained GMM system on NN input features. The Discrete Cosine Transform (DCT) follows to decorrelate Mel-filterbank features (FBANK). The speaker independent GMM-HMM system is done by single-pass retraining using these FBANKs. Finally, Constrained Maximum Likelihood Linear Regression (CMLLR) transform is estimated for each speaker.

We trained systems in iterative manner. In the first iteration, the simple monophone model was trained to get alignment of the text to the audio. In the second iteration, we got the final full system.

3.2 BLSTM

The ASR system developed for Babel 2016 [2] focuses on model training using BLSTM networks. The BLSTM system does not overperform the classical architecture but is more stable during training. The BLSTM network architecture consists of 3 hidden layers in both directions where there are 512 memory units in each layer and 300 neurons in the projection layer.

The transcriptions were not cleaned and were taken as is to train this system. The system was created in Kaldi and it is denoted as *Babel Kaldi BLSTM* in Table 1.

4. SUBWORD SYSTEM

The acoustic unit discovery (AUD) model presented in [4] aims at segmenting and clustering unlabeled speech data into phone-like categories. It is similar to a phone-loop model in which each phone-like unit is modeled by an HMM. This phone-loop model is fully Bayesian in the sense that:

- it incorporates a prior distribution over the parameters of the HMMs

- it has a prior distribution over the units modeled by a Dirichlet process [5].

Informally, the Dirichlet process prior can be seen as a standard Dirichlet distribution prior for a Bayesian mixture with an infinite number of components. However, we assume that our N data samples have been generated with only M components ($M \leq N$) from the infinite mixture. Hence, the model is no longer restricted to have a fixed number of components but instead can learn its complexity (i.e. number of units used M) according to the training data. The priors over the GMM weights, Gaussian mean and (diagonal) covariance matrix are a Dirichlet and a Normal-Gamma density respectively and were initialized as described in [6]. See [4] for the Variational Bayesian treatment of this model.

5. CONCLUSION

The primary on-time system based on BLSTM using original 16kHz audio and trained on the original transcriptions resulted in overall 48 % WER for the test set data. For ELSA test subset, the WER reached 4.9 % which is particularly good result. This subset data probably fits perfectly to the training set. On the contrary, the unseen YouTube test subset resulted in 87.2 % WER which is the worst score out of test subsets.

The improved late submitted BLSTM system using the combination of LMs showed overall 1.7 % WER improvement on the test data. The improvement on every single test subsets is nearly 3% WER (except ELSA subset).

The late GMM/DNN system using 8kHz audio and cleaned texts ended up with the worst overall score 55.7 % WER for the test set. Compared to our best BLSTM system, the score decreased by overall 9.4 % WER. The results for single databases (seen during the training) are following: the decrease of 23.4 % for ELSA; the decrease of 6.7 % for Forvo; the decrease of 12.6 % for RhinoSpike. The only score improvement of 3.3 % was for YouTube subset. The conclusion is that GMM/DNN system is more robust on unseen data.

6. REFERENCES

[1] Martin Karafiát, František Grézl, Mirko Hannemann, and Jan Černocký. BUT Neural Network Features for Spontaneous Vietnamese in BABEL. In *Proceedings of ICASSP 2014*, pages 5659–5663. IEEE Signal Processing Society, 2014.

[2] Martin Karafiát, Murali Karthick Baskar, Pavel Matějka, Karel Veselý, František Grézl, and Jan "Honza" Černocký. Multilingual BLSTM and Speaker-Specific Vector Adaptation in 2016 BUT Babel System. *Accepted at SLT 2016*, 2016.

[3] Igor Szöke and Xavier Anguera. Zero-Cost Speech Recognition Task at Mediaeval 2016. In *Working Notes Proceedings of the Mediaeval 2016 Workshop*, Hilversum, Netherlands, October 20-21 2016.

[4] Lucas Ondel, Lukáš Burget, and Jan Černocký. Variational Inference for Acoustic Unit Discovery. In *Procedia Computer Science*, volume 2016, pages 80–86. Elsevier Science, 2016.

[5] Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Annals of Statistics*, 2(6), November 1974.

[6] Chia-ying Lee and James Glass. A Nonparametric Bayesian Approach to Acoustic Model Discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 40–49, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.