ININ submission to Zero Cost ASR task at MediaEval 2016

Tejas Godambe, Naresh Kumar, Pavan Kumar, Veera Raghavendra, Aravind Ganapathiraju

Interactive Intelligence India Private Limited
Hyderabad, India

{tejas.godambe, naresh.kumar, pavan.kumar, veera.raghavendra, aravind.ganapathiraju}@inin.com

ABSTRACT

This paper details the experiments conducted to train an as good performing Vietnamese speech recognition system as possible using public domain data only, as a part of the Zero Cost task at MediEval 2016. We explored techniques related to audio pre-processing, use of speaker's pitch information, data perturbation, for building subspace Gaussian mixture acoustic model which is known for estimating robust parameters when the amount of data is less, and also unsupervised adaptation, RNN language model based lattice rescoring and system combination using ROVER tec hnique.

1. INTRODUCTION

The goal of the zero cost ASR task is to bring researchers together on the topic of training ASR systems using only data available in the public domain. In particular, this year's task consisted of the development of an LVCSR for Vietnamese language which is a rare enough language but with sufficient enough public data to work with. More details on this task can be found in [1].

Section 2 outlines the steps followed for building the final system. Section 3 describes in detail each experiment we conducted, and also discusses the loss/gain achieved in accuracy with it. We conclude the paper in Section 4.

2. APPROACH

We used the Kaldi ASR toolkit [2] for building the system. As lexicon was not provided, graphemes were used as phonemes. There were 96 unique phonemes. The below steps were followed for the development of the final system.

1. Truncate long silences in training data to 0.3 sec.
2. Augment data with speed perturbed versions (of speed factors 0.9 and 1.1) of itself [3].
3. Extract MFCCs along with pitch information [4].
4. Build SGMM acoustic model [5].
5. Construct a 5-gram language model (LM) from the training text.
6. Perform unsupervised adaptation, i.e. decode test utterances with above system, and add them to the training data along with their approximate hypothesized transcriptions. Three copies of test data (of speed factors 0.9, 1.0, 1.1) were added.
7. Generate lattices and rescore them with RNN based language model [6].
8. Do final decoding

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

3. RESULTS AND DISCUSSION

3.1 Preliminary Analysis

The sequence of experiments performed, and the gains/loss incurred in WER with each of them are detailed below. Table 1 shows the WER and the word error rate reduction (WERR) achieved for each individual experiment. The WER was calculated on a very small dev local data set which comprised of 21 utterances only.

1. **Using tri-phone model:** We first trained the tri-phone model with 2000 senones and total 20k Gaussians to see whether we are able to replicate the baseline result. This gave a WER of 37.0%
2. **Truncating silence in training data:** Preliminary observation of a few wave files showed presence of long silences, which usually corrupts the acoustic model. A WERR of 9.6% was achieved when the tri-phone model was trained after truncating long silences to 0.3 sec in the training data. Henceforth, for all experiments, we used the training data with truncated silences. This also reduced the size of the training data from around 13 hours to around 7 hours.
3. **Truncating silence in test data**: Inspired by the above gain, we truncated long silences to 0.3 sec in the test data too, before decoding. But, surprisingly, this increased the WER to 50.3%. Hence, in the future experiments, truncating silences in the test data was avoided.
4. **Using SGMM model:** SGMM model is known to estimate robust parameters and perform better than a simple tri-phone model, especially when the size of training data is small. A WERR of 9.4% was achieved upon migrating from tri-phone model to SGMM.
5. **Using DNN model:** DNNs are the state-of-the-art. But, it has been observed that they yield poorer or comparable results to SGMM when the size of training data is of small. We trained a basic DNN containing 429 nodes in the input layer (5 context frames), three hidden layers 512:256:512 with 256 being the bottleneck layer, and containing 930 output nodes, optimized using stochastic gradient descent to minimize the cross-entropy. But, this increased the WER to 23.5%. Though DNNs could have been made to perform better than SGMMs using proper regularization, because of time constraints, we stuck to the SGMM acoustic model.
6. **Using position independent phones:** This experiment was to see how the use of position independent phones fares against using position-dependent phones. Not so surprisingly, this step degraded the WER by 1%. So, position-dependent phones were used for further experiments.

7. **Unsupervised adaptation:** In unsupervised adaptation, we folded in the test data comprising of 332 utterances with their approximate hypotheses (obtained by decoding with SGMM in the previous run) into the training data, and re-trained the SGMM acoustic model. This gave 2.0% WERR.
8. **Audio augmentation 1:** Inspired by [3], speed of the original training data was perturbed by factors of 0.9 and 1.1, and these perturbed copies were augmented to the original training data. This helped achieve 1.1% WERR.
9. **Audio augmentation 2:** Here, four perturbed copies of speed factors 0.8, 0.9, 1.1 and 1.2 were augmented to the original training data. This gave 0.8% WERR, which is less than 1.1% achieved in the previous experiment. Hence, for the final system, we augmented original data with perturbed copies of speed factors 0.9 and 1.1 only.
10. **Using pitch information:** The confusing words in the hypothesis seemed to be acoustically close as many confusing pairs differed by just one phone. For some words, it appeared that the confusions are occurring because of different tonal manifestation of the same phone. This gave the idea of using pitch information along with traditional MFCCs as explained in [4]. This gave 1.2% WERR, and helped to eliminate a few recurring confusions.
11. **Using 5 gram LM:** Next, higher order N-grams were tried in order to put more constraints on the hypothesis and consequently improve the WER. Use of 5-gram LM instead of trigram LM helped achieve 2.0% WERR.
12. **Using 7 gram LM:** Inspired by the above gain, even higher order N-gram such as 7 grams were experimented. This gave 1.5% WERR which is less than 2.0% achieved with 5 grams. Hence, in the final system, 5-gram LM was used.
13. **Combined system:** For the final system we combined all the things that improvement such as truncating silence in the training data, using SGMM, unsupervised adaptation, data augmentation with speed factors 0.9, and 1.1, using pitch and using 5 gram LM. This combined system gave WER=13.8%.
14. **Rescoring lattices using RNN-LM:** Motivation behind using RNN LM [6] was to see how much gain we can achieve by putting more constraints (apart from the 5-gram LM) from the LM side using a model which captures long-term dependencies in text in a distinct manner than that done by N-grams. The lattices were re-scored using RNN LM, but it gave only 0.3% improvement. Probably limited amount of training text prevented getting full advantage of RNN-LM.
15. **Hypothesis combination;** ROVER [7] is a well-known technique to combine hypotheses from multiple different systems. Individual systems which had given improvements were combined with the above discussed combined system, but this did not yield better results than the combined system.

3.2 Final Results

In total, the test data comprised of 332 utterances, which contained utterances from ELSA, forvo.com, rhinospike.com and youtube.com. The percent WER achieved by our system on the above individual test data sets in the respective order are 5.7, 72.5, 25.3 and 91.4. The average WER is 51.2. While our system did well on data from ELSA and rhinospike.com, it did relatively poor on data from forvo.com and youtube.com.

Table 1: Sequence of experiments performed with individual WER and WERR

Row no.	Experiment	WER (%)	WERR (%)
1	Training the tri-phone model	37.0	
2	Truncating silence in training data	27.4	37.0-27.4=9.6
3	Truncating silence in test data	50.3	27.4-50.3=-22.9
4	Using SGMM model	18.1	27.4-18.1=9.3
5	Using DNN model	23.5	18.1-23.5=-5.4
6	Using position independent phones	19.1	18.1-19.1=-1.0
7	Unsupervised adaptation	16.1	18.1-16.1=2.0
8	Audio Augmentation 1	17.0	18.1-17.0=1.1
9	Audio Augmentation 2	17.3	18.1-17.3=0.8
10	Using pitch information	16.9	18.1-16.9=1.2
11	Using 5 gram LM	16.1	18.1-16.1=2.0
12	Using 7 gram LM	16.6	18.1-16.6=1.5
13	Combined system	13.8	
14	Rescoring lattices using RNN LM	13.5	13.8-13.5.3
15	ROVER	13.5	13.5-13.5=0.0

4. CONCLUSION

In this task, we confronted a real-world problem of building ASR system from public domain data containing noises and having imperfect transcripts. The data was inherently small in size. So, the problem of noisy acoustics and imperfect transcripts was multiplied with that of low-resource one. In our system, we tried to look at different aspects of ASR system building like audio pre-processing, data perturbation, using pitch information, acoustic modeling, language modeling using higher order N-grams, unsupervised adaptation, lattice-rescoring, and system combination. Each of the above techniques contributed their share toward bringing down the WER of the final system.

5. ACKNOWLEDGEMENTS

We thank the event and task organizers for their prompt responses to our queries related to the task.

6. REFERENCES

[1] Szoke, I., Anguera, X., 2016, Zero cost speech recognition task at MediaEval 2016, *In Working Notes, Proceedings of the MediaEval 2016 Workshop, Hilversum, Neth*erlands, 20-21 Oct 2016

[2] Povey, et al. 2011. The Kaldi speech recognition toolkit. *In Proceedings of ASRU*, 2011.

[3] Ko, Tom, et al. Audio augmentation for speech recognition *Proceedings of INTERSPEECH*. 2015.

[4] Ghahremani, Pegah, et al. A pitch extraction algorithm tuned for automatic speech recognition." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[5] Povey, Daniel, et al. "Subspace Gaussian mixture models for speech recognition." *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010.

[6] Mikolov, Tomas, et al. "Rnnlm-recurrent neural network language modeling toolkit." *Proc. of the 2011 ASRU Workshop*. 2011.

[7] Fiscus, Jonathan G. "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)." *Automatic Speech Recognition and Understanding, 1997. Proceedings. 1997 IEEE Workshop on*. IEEE, 1997