Efficient Acoustic Modeling Method for Unsupervised Speech Recognition using Multi-Task Deep Neural Network

Yao Haitao[1, a], An Maobo[2, b], Xu Ji[1], Liu Jian[1]

[1]Institute of Acoustics, Chinese Academy of Sciences, No. 21 North 4th Ring Road, Haidian District, 100190 Beijing, China

[2] National Computer network Emergency Response Technical Team/Coordination Center of China, Beijing, 100029

[a]email: yaohaitao@hccl.ioa.ac.cn, [b]email: amb@cert.org.cn

Keywords: Speech Recognition, Acoustic Modeling, Unsupervised Training, Multi-Lingual, Multi-Task Deep Neural Network

Abstract. This paper proposes a method of acoustic modeling for zero-resourced languages speech recognition under mismatch conditions. In those languages, very limited or no transcribed speech is available for traditional monolingual speech recognition. Conventional methods such as IPA based universal acoustic modeling has been proved to be effective under matched acoustic conditions (similar speaking styles, adjacent languages, etc.), while usually poorly preformed when mismatch occurs. Since mismatch problems between languages often appears, in this paper, unsupervised acoustic modeling via cross-lingual knowledge sharing has thus been proposed: first, initial acoustic models (AM) for a target zero-resourced language are trained using Multi-Task Deep Neural Networks (MDNN) – different languages' speech mapped to the phonemes of the target language (mapped data) is jointly trained together with the same data transcribed language specifically and respectively (specific data); then, automatically transcribed target language data is used in the iterative process to train new AMs, with various auxiliary tasks. Experiment on 100 hour Japanese speech without transcripts achieved a character error rate (CER) of 57.21%, 19.32% absolute improvement compared to baseline (IPA based universal acoustic modeling).

Introduction

Recently, automatic speech recognition (ASR) for low resourced languages has become an active research field [1]. Traditional mono-lingual ASR systems require a certain amount of transcribed speech data [2]. Since it is not easy (expensive, time-consuming, lack of linguistic experts) to collect enough manually transcribed speech data of new low or even zero resourced languages, it is a big challenge to build recognition systems for these languages. Acoustic modeling, which required a large amount of accurate transcribed speech, is one of the toughest tasks among low-resource speech recognition.

Unsupervised or semi-supervised training using multilingual information is a promising means to train AM with low cost. As described in [3], rapid development of an automatic speech recognition system can greatly benefit from the use of unsupervised acoustic model training. Initial models, or so called seed models, are used to generate transcripts for unsupervised speech [4]. Then, data selection [5] [6] (confidence based or majority voting, etc.) is adopted to get more reliable training speech. After that, training process might be applied to improve the recognition performance iteratively [7] [8].

In semi-supervised scene, shared hidden layer multi-softmax deep neural network [9] which is jointly trained of supervised and automatically transcript unsupervised data proves to work well [10].Whereas, in completely unsupervised scenario, the problem is that, no transcribed data is available and it is often hard to build effective seed models.

An alternative method to build seed model for an unsupervised language is to do phoneme mapping, either all languages use a universal phoneme set (for example, IPA – International Phonetic Alphabet) or some languages' phonemes are mapped to others'. Vu et al. demonstrated the

effectiveness of phoneme mapping under matched conditions, from Czech to Bulgarian, Croatian, Polish, Russian in accordance with IPA [11] [12] similarities. But whether this method works under mismatch conditions still remains to be seen. Iterative process is adopted to deal with mismatch conditions. In these previous studies, the problem of transcript errors, which might be even harmful [10], is serious and has not been solved, especially at the early stage of unsupervised training.

As introduced above, mismatch conditions and transcription errors are the two main problems that multi-lingual unsupervised training involves. In this paper, a new method involves MDNN in both initial and iterative training process is proposed to solve these two problems. The remainder of this paper is organized as follows. In section 2, we describe our proposed method in detail. Section 3 presents data resources we use and the baseline system. Section 5 reports the experimental setup and results. The study is concluded in section 5 with a summary and an outlook to future steps.

Proposed Multi-Lingual Unsupervised Training Method

The multi-lingual unsupervised training method proposed here is aimed to handle the mismatch conditions and transcript errors. Three main parts are contained: a) Using language specific data as complementary tasks for mapped data, to lower the influence of mismatched data mapping from source language to target language, and thus, produce more reliable seed models; b) Acoustic stability [11] is used to take the place of confidence based score in the data selection of unsupervised training, since confidence scores generated by poorly estimated acoustic models do not perform well [13]; c) Iterative process: completely new models trained by MDNN, automatically transcribed data for primary task (task that keeps at last), multi-lingual and CI information for complementary tasks.

Firstly, we give a brief introduction of MDNN. Figure 1 shows the typical structure. MDNN is a multi-task learning (MTL) technique [14] that improves single-task learning (STL) by training the deep neural network (DNN) with several related tasks (each task is an output layer) and some shared hidden layers. These secondary tasks are used for the training stage and are dropped in the end.

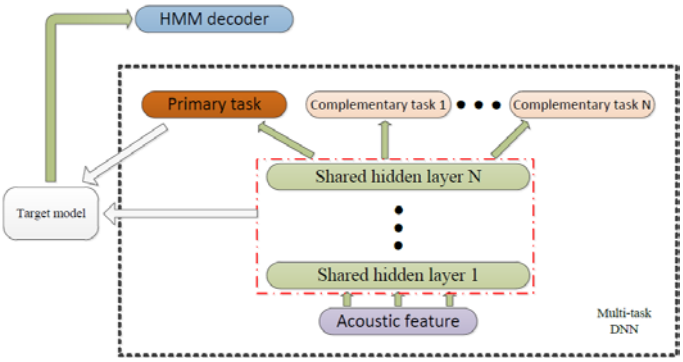

Fig.1. Typical structure of MDNN

Negative cross-entropy is defined as the objective function of STL-DNN:

$$F_A(\theta) = \sum_t \log p(y_t^A \mid x_t; \theta) \tag{1}$$

The parameter matrix θ is optimized with stochastic gradient descent to maximize this objective function. The gradient is computed with respect to small mini-batches of training frames, θ is updated with a small step size. Some epochs is carried out, each with several iterations to cover the complete training set once.

While in MDNN, new label of training data is prepared for additional task *B, C,* $\cdots$ with $F_B(\theta)$, $F_C(\theta)$, $\cdots$ being the objective function:

$$F(\theta) = \sum_i \lambda_i F_{T_i}(\theta), (\sum_i \lambda_i = 1) \tag{2}$$

All parameters except those in the output layers are shared across tasks.

MDNN is use in the initial step as follows. First, phoneme mapping is done from source language to the target. In this step, two kinds of mapping is adopted: one is forced mapping, which

means all source languages' phones are forced to map to the closest target language phone; the other is natural mapping, in which IPA phoneme sets are used for both source and target languages. We use phoneme-level lexicons and transcripts and only source language phonemes that can be seen in the target language's phoneme set are used as useful training data, other phones that are not appeared in the target language are seen as OOVs and wiped off. Then, the mapped data is used to simulate "target language data" and update the primary task of MDNN while the original source data is used to train the complementary tasks.

In the training step, CI information is added as new complementary tasks for context dependent (CD) primary tasks. As CI has proved to be of higher frame accuracy and has more training frames for each out senones, especially when training data is limited [15]. Also, it enriches the training data and brings down the effect of transcription errors and mismatch problems by joint updating the share hidden layers.

Acoustics stability feature (A-stab) [11] is used in the data selecting step. To compute this feature, a number of alternative hypotheses with different weighting between acoustic scores and language model (LM) scores is computed. Both forced mapping AM and natural mapping AM is used. Each of these hypotheses is aligned against the reference output of the recognition, where the reference output is defined as the output with the assumedly best weighting between AM and LM. Here we use natural mapping AM (since it give more accurate information) with LM scale 11, which proves to be best for the most part by experience.

For each word of the reference output, the A-stab score is defined as the number of times the same word occurs in the set of alternative hypotheses, normalized by the number of alternative hypotheses:

$$Word : S_w = \frac{\#num_{occur}(word_{ref})}{\#num_{hyp}} \tag{3}$$

For each sentences of the reference output, the A-stab score is defined as average of word scores:

$$Sentence : S_s = \frac{\sum_{\#num_{word}} S_w}{\#num_{word}} \tag{4}$$

Sentences with score higher than a given threshold are chosen as usable reference.

Data Resources and baseline system

The four languages Mandarin, English, Korean and Japanese in our experiment belong to various language families or branches. Japanese is the target language while the others act as source languages. Table 1 shows the phoneme distributions of Mandarin, English, Korean and their IPA phoneme coverage of Japanese. Mismatch problem exists between source and target languages, obviously.

Tab.1. Phoneme distributions

Languages	Language Specific Phoneme Number	IPA Phonemes Number	IPA Phonemes Covers Japanese
Mandarin	66	41	12
English	39	39	18
Korean	66	21	15
Total		66	24
Japanese	40	46	26

The source languages data we use are as follows: for Mandarin, we use a 100 hours split of LDC2005S15 HKUST Mandarin Telephone Speech; for English, 100 hours data from part 1 of fisher dataset is selected; for Korean, we use 100 hours of our self-collected speech. We experimentally evaluated the performance of our proposed method on 100 hours Japanese

unsupervised data, which is also self-collected. All speech are Conversational Telephone Speech (CTS) in 8 KHz 16 bit PCM format. Lexicon size is as follows: 45k words for Mandarin, 63k for English, 20k for Korean and 55k for Japanese. LM for Japanese is: 50k 1-gram items, 3544k 2-gram items, 10609 3-gram items. We did not necessarily need LMs of Mandarin, English and Korean.

The baseline system consists of two parts: initial step and iterative step. In the initial step, we do IPA based universal acoustic modeling using Mandarin, English and Korean speech. Then, traditional unsupervised iterative process is adopted for Japanese.

The Kaldi toolkit [16] is used for speech recognition framework. Standard 52-dim PLP feature (13-dim together with its 3 deltas), is extracted and used for maximum likelihood GMM model training. After that, a DNN-HMM hybrid system is trained using the 52-dim PLP as input and GMM-aligned senones as targets. For DNN, an 11-frame window is used in the input layer, we use 6 hidden layers, each has 2500-250 p-norm neuron with p=2 [17].

Mini-batch SGD is used for back propagation. The training starts with an initial learning rate of 0.008 and ends with a final learning rate of 0.0008 after 10 epochs.

We train Japanese AM on the 100 hour Japanese with its real transcripts to get a result of 42.23% CER, which serves as the upper bound in our experiment.

Experiments of Proposed Method

We first evaluate the performance of initial steps. In the proposed method, as described above, we use two initial AMs. One is based on the forced mapping, with 40 Japanese phonemes and about 5000 tri-phone states, lexicons of the three source languages are all in Japanese phonemes and merged during training process; the other is based on natural mapping, with 26 Japanese IPA phonemes and about 5000 tri-phone states, phoneme lexicons are used so that only the 24 Japanese are treated as Japanese In-Vocabulary word. 2 unseen Japanese phonemes are replaced by nearest phonemes in source languages. Input and hidden layer config of MDNN is the same as baseline. Table 2 gives out CER of the two proposed initial AMs and the baseline initial AM. We evaluate performance of different amount of mapped data to validate the effectiveness of shared hidden layers.

Tab.2. Performance of the Initial Steps

Model	Training Method	CER (%)
Baseline IPA Model	DNN	76.53
Forced Mapping Model (300h mapped data)	DNN	73.05
Forced Mapping Model (300h mapped data)	MDNN	71.88
Forced Mapping Model (100h mapped data)	MDNN	71.16
Natural Mapping Model (300h mapped data)	DNN	72.84
Natural Mapping Model (300h mapped data)	MDNN	71.20
Natural Mapping Model (100h mapped data)	MDNN	**69.83**

As shown above, 6.7% absolute CER decrease is got, using the proposed initial method. Two conclusions are drawn: MDNN is useful during initial model training, this is due to usage of multi-lingual information and that MDNN lightens the effect of errors in data mapping; a bit smaller amount of mapped data is better, since less effect it has on the hidden layer of MDNN.

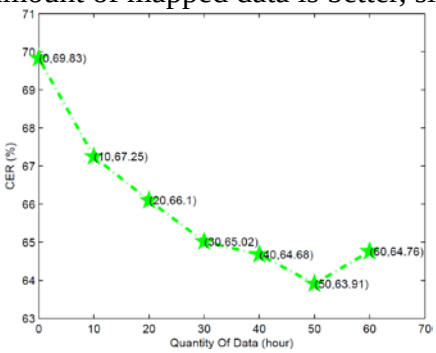

Fig.2. Quantity of Data and CER, iteration 1

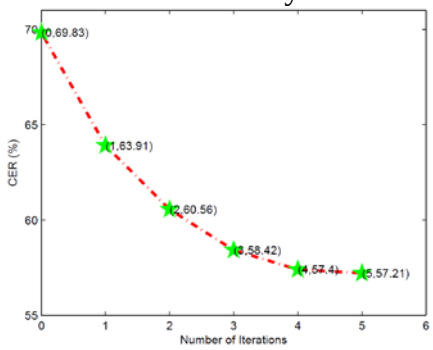

Fig.3. CER performance of Iterative Process

Then, we experiment to find how quantity of selected data affect performance of the training process in the first iteration after initial models. As shown in figure 2, we can see that 60 hours of data got the best performance in the training process, iteration 1. There exists a trade-off between data quality and quantity. In the following experiments, we use fixed amount of training data, 60 hours.

We give out performance of each iteration in our proposed method in figure 3. We can see that CER of Japanese decreases obvious, especially in the early stage of the iterative process. We get 12.62% absolute CER decrease compared to the initial step.

Conclusion

In this paper, we apply shared hidden layer MDNN in our unsupervised AM training. Data mapping is used to simulate training data for the unsupervised language. Although mismatch problem appears, we can get initial AM for the unsupervised language in its own phoneme and lexicon resource. Iterative process of training can deal with the mismatch conditions and get better performance. With 100 hours unsupervised data, our proposed method achieved CER of 57.21%, 19.32% absolute improvement compared to baseline. This result demonstrates the possibility of building AMs inexpensively. It would be interesting to investigate whether adaptation method can be used in the iterative process to interact with training process and benefit from each other. There still remains 14.98% absolute CER improvement to reach the upper bound of unsupervised training.

Acknowledgement

This work is partially supported by the National 863 Program (2015AA016306), the National 973 Program (2013CB329302), the National Natural Science Foundation of China (Nos. 11461141004), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500) and the National 863 Program (No. 2012AA012503).

References

[1] Besacier L, Barnard E, Karpov A, et al. Automatic speech recognition for under-resourced languages: A survey [J]. Speech Communication, 2014, 56: 85-100.

[2] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups [J]. Signal Processing Magazine, IEEE, 2012, 29(6): 82-97.

[3] Zavaliagkos G, Colthurst T. Utilizing untranscribed training data to improve performance [C]//DARPA Broadcast News Transcription and Understanding Workshop. 1998: 301-305.

[4] Grézl F, Karafiát M. Semi-supervised bootstrapping approach for neural network feature extractor training [C]//Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013: 470-475.

[5] Kemp T, Schaaf T. Estimating confidence using word lattices [C]//EuroSpeech. 1997.

[6] Wei K, Liu Y, Kirchhoff K, et al. Submodular subset selection for large-scale speech training data [C]//Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 3311-3315.

[7] Lööf J, Gollan C, Ney H. Cross-language bootstrapping for unsupervised acoustic model training: rapid development of a Polish speech recognition system [C]//Interspeech. 2009: 88-91.

[8] Lamel L, Gauvain J L, Adda G. Unsupervised acoustic model training [C]//Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on. IEEE, 2002, 1: I-877-I-880.

[9] Huang J T, Li J, Yu D, et al. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers [C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 7304-7308.

[10] Su12 H, Xu H. Multi-softmax Deep Neural Network for Semi-supervised Training [J]. 2015.

[11] Vu N T, Kraus F, Schultz T. Multilingual A-stabil: A new confidence score for multilingual unsupervised training [C]//Spoken Language Technology Workshop (SLT), 2010 IEEE. IEEE, 2010: 183-188.

[12] Vu N T, Kraus F, Schultz T. Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil [C]//Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011: 5000-5003.

[13] Saiko M, Yamamoto H, Isotani R, et al. Efficient multi-lingual unsupervised acoustic model training under mismatch conditions [C]//Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014: 24-29.

[14] Caruana R. Multitask Learning: A Knowledge-Based Source of Inductive Bias1 [C]//Proceedings of the Tenth International Conference on Machine Learning. 41-48.

[15] Bell P, Renals S. Complementary tasks for context-dependent deep neural network acoustic models [C]//Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[16] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit [J]. 2011.

[17] Zhang X, Trmal J, Povey D, et al. Improving deep neural network acoustic models using generalized maxout networks [C]//Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 215-219.