Time Delay Neural Network for Myanmar Automatic Speech Recognition

Myat Aye Aye Aung
Natural Language Processing Lab
University of Computer Studies,Yangon
Yangon, Myanmar
myatayeayeaung@ucsy.edu.mm

Win Pa Pa
Natural Language Processing Lab
University of Computer Studies,Yangon
Yangon, Myanmar
winpapa@ucsy.edu.mm

Abstract

Time Delay Neural Network (TDNN) contains in neural network architectures. In Automatic Speech Recognition, TDNN is strong possibility in context modeling and recognizes phonemes and acoustic features, independent of position in time. There are many techniques have been applied for improving Myanmar speech processing. TDNN based acoustic model for Myanmar ASR in this paper. Myanmar language is a low resource language and no pre-collected data is available. A larger dataset and lexicon than our previous work are applied in this experiment. The speech corpus contains three domains: Names, Web News data and Daily conversational data. The size of the corpus is 77 Hrs and 2 Mins and 11 Secs and include 233 female speakers and 97 male speakers. The performance of TDNN for Myanmar ASR is shown by comparing with Gaussian Mixture Model (GMM) as a baseline system, Deep Neural Network (DNN) and Convolutional Neural Network (CNN). Experiments evaluation is used 2 test data: TestSet1, web news and TestSet2, recorded conversational data. The experimental results show that TDNN outperforms GMM-HMM, DNN and CNN.

Keywords: *GMM-HMM, DNN, CNN, TDNN, acoustic modelling*

I. INTRODUCTION

Automatic Speech Recognition (ASR) aims to enable computers to "understand" human speech and convert it into text. ASR is the next frontier in intelligent human-machine interaction and also a precondition for perfecting machine translation and natural language understanding. It contains two models; the language model models probabilities of word sequences and the acoustic model describes distributions of acoustic features for individual phones. Typically, the two statistical models are independently trained from large volumes of text data and annotated speech data, respectively. The component connecting these two models is the pronunciation lexicon mapping words into phone sequences [1].

ASR requires an acoustic model that can effectively learn the context of adjacent input speech features to improve the recognition performance [2]. Nowadays, almost all many speech recognition systems have been used neural networks to achieve recognition performance.

Neural Networks are requires many speech data in convergence of model training, and as a result, it consumes a lot of time in model training. Time-delay neural network (TDNN) [3] is utilized to model long-term dependencies. Sequence classification is transformed to multidimensional curve classification by TDNN.

Neural network architectures have been applied to advantage for speaker adaptation. iVectors is used to extract information about speaker or acoustic environment, which have been exposed to be valuable for instantaneous and discriminative adaptation of the neural network [4]. iVectors-TDNN based acoustic model is applied in this work to improve the performance of Myanmar ASR and the results are evaluated using our own Myanmar speech corpora, UCSY-SC1 [5].

There are some recent works on Myanmar ASR using different Machine Learning approaches. H. M. S. Naing, et. al., [6] presented a large vocabulary continuous speech recognition on Myanmar language, applying three acoustic models of One Gaussian Mixture Model (GMM) and two Deep Neural Networks (DNNs) on 40 hours of speech dataset. The evaluations were done on an open test set of 100 utterances, recorded by 25 Native speakers. Word Error Rate (WER) reached up to 15.63 % in the sequence discriminative training DNN.

A. N. Mon et. al., presented CNN based Myanmar ASR, building Myanmar speech corpus, 20 hours read speech recorded by 126 females speakers and 52 males speakers, and evaluated with two datasets, opened-data, web news data and closed-data, recorded data, and achieved 24.73% and 22.95% of Word-Error-Rates[7].

A speech corpus, UCSY-SC1, is introduced in [5], and it is evaluated by comparing GMM-HMM, DNN and CNN, for improving Myanmar ASR, showing their experimental results, 15.61% and 24.43% (WER).

TDNN-based acoustic model is applied for Korean corpora [8] shows has an advantage in fast-convergence on TDNN when the size of training data is restricted, as sub-sampling excludes duplicated weights. and not as an independent document.

II. AUTOMATIC SPEECH RECOGNITION

Main Component of Automatic Speech Recognition system are Pre-processing, Feature Extractor, Acoustic Model, Language Model and Recognizer. An acoustic model's task is to compute the P (O|W), i.e. the probability of generating a speech waveform for the model [9].

A. GMM-HMM Model

GMM is a probabilistic model which is signified as a biased amount of Gaussian element densities and used to model the distribution of the acoustic characteristics of speech. Gaussian distribution is intended by mean, variance and weight. HMM is used to represent the transition probabilities between states.

The transition between phones and corresponding observable can be modeled with the Hidden Markov Model (HMM). An HMM model composes of hidden variables and observables [10]. The horizontal arrows demonstrate the transition in the phone sequence. As shown in Fig. 1 [16].

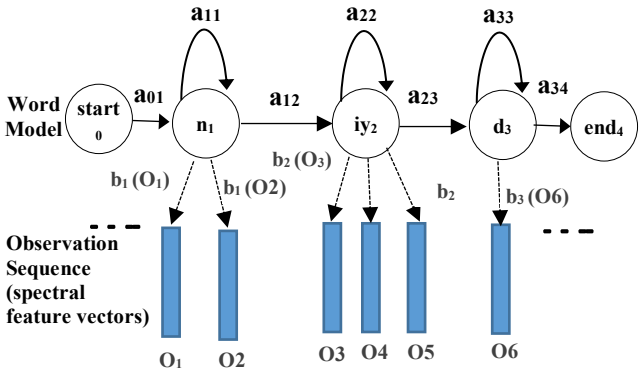

Figure 1. Typical HMM architecture

GMM-HMM system, 39 features of employed for training and testing of this system using MFCC feature extraction technique. It contains 13 static energy of first and second changing derivations order delta features and cepstral mean and variant normalization. The implementation is done on 25 and 10 ms frame length and with a frame shift of 10ms is sliced from 9 frames with combination and project down into 40 feature vectors, followed by applying linear discriminant analysis (LDA) to lower 40 dimensions. The predictable vectors are handled with maximum likelihood linear transform (MLLT) method to produce de-correlation. The feature space maximum likelihood linear regression (fMLLR) is applied in this work for speaker adaptive training (SAT). The preparation of these vectors are combined with GMM-HMM classifiers. The training is started with state of a decision tree for tri-phone model and mono-phone model.

B. Deep Neural Network (DNN)

Deep Neural Networks are many hidden layers of units between input and output layers. They are typically feed forward neural networks (FFNNs). DNNs contain many hidden layers with a large number of non-linear units and a large output layer. The large output layer is performance as an input to various number of HMM states. DNN-HMM outperformed traditional GMM-HMM due to increased modeling power [11].

For DNN-HMM based model, feature space MLLR (fMLLR) method is applied in this experiment. Mel-Filter Bank features estimated in speaker adaptive training approach and is worked on GMM based system. Four hidden layers with three hundred units per hidden layers used in this work. The process is started high dimension 1024. Same process of fMLLR transformation is done in development and testing phase which is applied in decoding module.

C. Convolutional Neural Network (CNN)

CNNs are popular models of deep learning that are widely used in ASR. They are decreasing spectral variations and modeling spectral correlations in acoustic features. Hybrid speech recognition systems including CNNs with HMM-GMM have accomplished the state-of-the-art. CNNs consists of weight sharing, convolutional filters, and pooling. Therefore, CNNs have reached an impressive performance in speech processing. CNNs are composed of many convolutional layers. Pooling is that decreases the dimensionality of a feature map [12]. The same setting of CNNs [5] is applied in this work for evaluation.

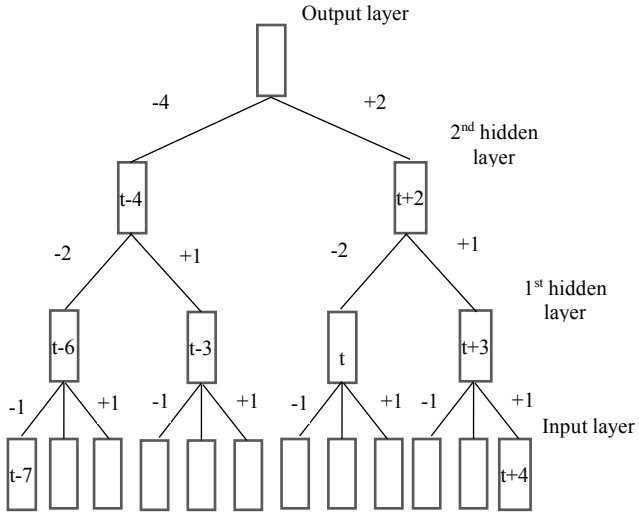

Figure 2. Typical structure of TDNN architecture

D. Time Delay Neural Network (TDNN)

Time-delay neural networks are designed to express a relation among inputs in time for speech recognition. TDNN design the initial transforms are learned within narrow temporal contexts while the later layer operate a wider temporal context [8]. Example of TDNN architecture is shown in Fig. 2 [17].

TDNN architecture are tied across time steps, which is to decrease parameters and learn shift invariant feature transforms. Another way is used to reduce parameters and advance calculation is sub-sampling, for example, the splicing configuration {-1, 1} means that splice the input at present period substract 1 and the current time step add 1 [2] [8]. Input context {-1, 1} is applied in this experiment. 100-dimensional iVector to the 43-dimensional input apply on each frame. 11 hidden layers and bottleneck-dimension is 256 and a constant learning rate is 0.001. The number of epochs 20 and mini-batch-size 128 is used in this work.

III. PRONUNCIATION LEXICON

The acoustic model uses phonemes and language model is output words, a lexicon is needed to connection the gap. The lexicon is used to translate words or sentences into a sequence of phonemes. The lexicon needs to able to map each word in the language model to a sequence of phonemes. A Grapheme-to-Phoneme (G2P) converter can automatically transcribe the words needed for the lexicon.

This experiment of lexicon was created by training a G2P conversion model using Myanmar Dictionary. The dictionary includes in total about phonetically transcribed words. This paper used a lexicon that has the word size of 44,376 that is the extension of Myanmar Language Commission (MLC) [13]. There are 110 phonemes in the training set.

IV. SPEECH DATASET

UCSY-SC1 [5] formed by combining data with web news and conversational data. This speech corpus contains three domain: Names, Web News and Daily conversational data. Names are unique 2,250 sentences and obtained from UCSY website[1], which is recorded with 10 intern students. Conversation data have been collected from UCSY NLP Lab members and Internship students with 58 speakers and have over 7,200 unique sentences. Names and Daily conversation are recorded with voice recorder Tascam DR-100MKIII[2]. Sample frequency used with 16 kHz. 832,658 words and 15,095 unique words in this corpus. Web News Data are from Myanmar News Websites such as MRTV news, Eleven News, For Info News, 7Day TV. TestSet1 and TestSet2 are Web News and conversation data. The statistics of corpus is shown in TABLE I.

TABLE I. STATISTICS OF SPEECH CORPUS

Data	Size	Speakers			Utterances
		Female	*Male*	*Total*	
Names	6Hrs 30Mins	6	4	10	22,250
Web News	25Hrs 20Mins	177	84	261	9,066
Daily Conversation	45Hrs 3Mins	42	4	46	72,003
TestSet1	31Mins 55Secs	5	3	8	193
TestSet2	32Mins 40Secs	3	2	5	887
Total	77Hrs 2Mins 11Secs	233	97	330	104,399

TABLE II. INPUT FEATURES

Acoustic model	No. of Hidden layers	Learning rate	No. of feature dimensions (mfcc)
DNN	4	0.008	40
CNN	4	0.008	40
TDNN	11	0.001	43

The details input features are used in this work for neural network experiments as shown in TABLE II.

V. EXPERIMENTAL SETUP

This experiment developed with 76 Hrs and 53 Mins training data. SRILM language modeling toolkit [14] is used to build language model. The details statistics are shown at TABLE III. The development dataset was taken from training set. Dataset contain two sets: TestSet1 and TestSet2, which are the same with UCSY-SC1 [5] corpus. TestSet1 is open test data, which is Web News. It is generated by 8 speakers. TestSet2 is also open test set with native 5 speakers recorded to the conversational data with recorders. For all the neural networks training is worked on TESLA K80 GPU. The system experiments are done using Kaldi [15] toolkit version for development.

The evaluate of speech recognition systems is measured word error rate (WER) [7], which it can then be computed as:

$$WER = \frac{100 \times (insertions + substitutions + Deletions)}{TotalWord \in CorrectTranscript} \quad (1)$$

[1] http://ucsy.edu.mm/

[2] https://tascam.com/us/product/dr-100mkiii/top

TABLE III. STATISTICS ON TRAIN AND TEST SETS

Data Setup	Size	Speakers			Utterances
		Female	*Male*	*Total*	
Training	76Hrs 53Mins	225	92	317	103,319
Developm-ent	30Mins 12Secs	6	4	10	213
TestSet1	31Mins 55Secs	5	3	8	193
TestSet2	32Mins 40Secs	3	2	5	887

TABLE IV. EXPERIMENTAL RESULT

Model	Dev (%WER)	TestSet1 (%WER)	TestSet2 (%WER)
	Speaker Independent	*Speaker Independent*	*Speaker Independent*
GMM-HMM	21.50	26.11	28.78
DNN	15.84	20.20	22.60
CNN	15.55	18.44	20.50
TDNN	**11.25**	**15.03**	**16.83**

VI. EXPERIMENTAL RESULT

Myanmar ASR performance is evaluated using TDNN acoustic modeling technique and compared with baseline GMM-HMM, DNN, CNN model as shown in TABLE IV.

TABLE IV shows word error rates of Development data (Dev), TestSet1 and TestSet2 based on training data. Four acoustic model are compared. This experiment use the same lexicon and same language model. TDNN is outperforming than 4 acoustic models in speaker independent.

VII. CONCLUSION

This paper presents the extension of speech corpus UCSY-SC1 [5] for Myanmar ASR. The large speech corpus is to support for Myanmar speech development because Myanmar is a low-resourced language. GMM-HMM, DNN, CNN and TDNN are applied to evaluate the performance of extended speech corpus. TestSet1 and TestSet2 are used to estimate the ASR accuracy. TDNN is better than base line acoustic model GMM-HMM, and DNN and CNN. Applying sub-sampling in TDNN decreases the model size that reduces the number of parameters for the hidden layers, so that the dimensions of hidden layers decreased significantly. TDNN also leads to the lowest error rates on both TestSet1 and TestSet2.

Building the speech corpora is important for low-resourced Myanmar language and expected that this corpus will be some of use for more Myanmar speech processing. End-to-End learning approach will be applied in our future work to improve the performance of Myanmar Automatic Speech Recognition.

REFERENCES

[1] A. Cloud, "Inter speech 2017 Series Acoustic Model for Speech Recognition Technology," Alibaba Clouder Blog.

[2] V. Peddinti, D. Povey, S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts", Proceedings of Interspeech, 2015, 3214–3218.

[3] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using Time Delay Neural Network," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328-339, Mar. 1989.

[4] S. Xue, O. Abdel-Hamid, H. Jiang, L. Dai, and Q. –F. Liu, " Fast Adaptation of Deep Neural Network based on Discriminant Codes for Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. PP, no. 99, pp. 1-1, 2014.

[5] A. N. Mon, W. P. Pa and Y. K. Thu, "UCSY-SC1: A Myanmar speech corpus for automatic speech recognition," International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 4, August 2019, pp. 3194_ 3202.

[6] H. M. S. Naing, A. M. Hlaing, W. P. Pa, X. Hu, Y. K. Thu, C. Hori, and H. Kawai, "A Myanmar Large Vocabulary Continuous Speech Recognition System", In Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2015, Hong Kong, December 16-19, 2015, pages 320–327, 2015.

[7] A. N. Mon, W. P. Pa, Y. K. Thu and Y. Sagisakaa, "Developing a speech corpus from web news for Myanmar (Burmese) language," 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, 2017, pp. 1-6.

[8] H. Park, D. Lee, M. Lim, Y. Kang, J. Oh and J. H. Kim, "A Fast- Converged Acoustic Modeling for Korean Speech Recognition: A Preleminary Study on Time Delay Neural Network", To cite this article: Boji Liu et. al. 2019 J.Phys : Conf. Ser. 1229 012078.

[9] Suman K. Saksamudre, P. P. Shrishrimal, R. R. Deshmukh, "A Review from Different Approaches for Speech Recognition System," International Journal of Computer Applications (0975-8887), Volume 115-No.22, April 2015.

[10] P. Bansal, A. Kant, S. Kumar, A. Sharda, S. Gupta, "IMPROVED HYBRID MODEL OF HMM/GMM FOR SPEECH RECOGNITION," International

Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008.

[11] Lekshmi.K.R, Dr.Elizabeth Sherly, "Automatic Speech Recognition using different Neural Network Architectures – A Survey," Lekshmi.K.R et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (6) , 2016, 2422-2427.

[12] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Gerald Penn, "APPLYING CONVOLUTIONAL NEURAL NETWORKS CONCEPTS TO HYBRID NN-HMM MODEL FOR SPEECH RECOGNITION," Conference Paper in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · May 2012.

[13] M.L.Commission, "Myanmar-English Dictionary," Department of the Myanmar Language Commission, Yangon, Ministry of Education, Myanmar, 1993.

[14] A.Stolcke, "Srilm - An Extensible Language Modeling Toolkit", pp. 901--904 (2002).

[15] D.Povey, et al., "The Kaldi Speech Recognition Toolkit," Idiap, 2011.

[16] P. Janos-Pal, "Gaussian Mixture Model and the EM algorithm in Speech Recognition," slideplayer.com.

[17] I. Kipyatkova, "Experimenting with Hybrid TDNN/HMM Acoustic Models for Russian Speech Recognition," Springer International Publishing AG 2017, A. Karpov et. al. (Eds): SPECOM 2017, LNAI 10458, pp. 362-369, 2017.