The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Evaluating Code-Switched Malay-English Speech Using Time Delay Neural Networks

Anand Singh, Tien-Ping Tan

School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia
anandsingh@student.usm.my, tienping@usm.my

Abstract

This paper presents a new baseline for Malay-English code-switched speech corpus; which is constructed using a factored form of time delay neural networks (TDNN-F), which reflected a significant relative percentage reduction of 28.07% in the word-error rate (WER), as compared to the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM). The presented results also confirm the effectiveness of time delay neural networks (TDNNs) for code-switched speech.

Index Terms: automatic speech recognition, code-switching, acoustic modeling, time delay neural networks, low-resourced languages

1. Introduction

From the last several decade's human beings are trying to develop speech technologies that could recognize speech perfectly. Humans can recognize the speech in a better manner because they make the appropriate use of acoustic, linguistic and contextual information; significant improvements are highly needed in the Automatic Speech Recognition (ASR) systems to reduce the human-machine performance gap [1]. Due to technological improvements in the field of speech recognition, ASR technology has shifted from the laboratory to commercial markets and is now widely used in mobile devices, like Apple's Siri and in many other realistic applications, like Google's voice search [2-5]. For ASR, the main goal is to automatically transcribe the speech signal in terms of a sequence of items as close as possible to a referenced transcription [6]. Conventional model, like Hidden Markov Models (HMMs) is performing well from several decades [7], a fusion of NNs and deep learning, was introduced a few years ago [8]. Collaboration between researchers from academia, like the University of Toronto and industrial research groups like Google research, IBM research and Microsoft research have been responsible for this tremendous growth after 2010 [9]. This growth also gets the support of the Graphics Processing Units (GPUs), and due to this increased computational power, several-year-old practices replaced by the powerful methods.

ASR systems consist of three models; namely; acoustic model, which characterizes the sound of the language, mainly all the phonemes understood by the ASR systems, acoustic models are built through a training operation using a large quantity of transcribed data. Lexicon model; contains the phonetic lexicon of the words that can be recognized by the system with their possible pronunciations and the language model; provides knowledge about the word sequence that can be uttered. In the state-of-the-art approaches, acoustic model, lexicon and language model decide the probable sentence given in a speech utterance [10-12]. In acoustic modeling, currently, more focus is on DNN-HMM or DNNs. Following Figure 1 shows, a typical ASR architecture.

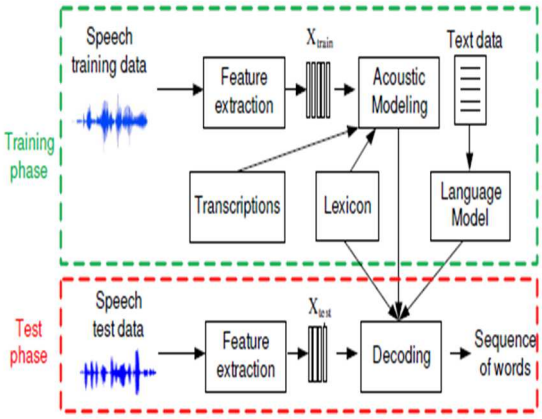

Figure 1: *An Overview of a typical ASR system [13]*

In this work, we presented our latest results on Malay-English code-switched speech recognition because Malaysian speakers often code-switch between Malay and English in daily conversations [14]; This is to address different audiences, to emphasize a point, various other social factors and plays an important role in education, tourism and business life [15-18].

Recently, TDNNs, have produced good improvements in many ASR tests [25]. So, it gives the motivation to use TDNNs in the low-resourced; code-switched data scenario.

The rest of this paper is organized as follows. Section 2 describes the basic acoustic models in greater detail. Section 3 elaborates the weighted finite state transducers. Section 4 is related with the experimental setup, under which database profile and training are explained. The results mentioned in section 5. At last, section 6 presents the conclusion and future scope of the work.

2. Acoustic models

2.1. GMM-HMM

GMM-HMM-based acoustic models were very popular because it can model acoustic variations of a speech signal very well even though the amount of training is small, and computationally it takes less time in training and testing. As HMM take care of the temporal variability of speech and GMM plays a vital role in capturing the spatial variations of the speech; can be described as parametric probability density function (pdf); that can be presented as a weighted sum of

10.21437/SLTU.2018-41

Gaussian component densities as mentioned by the equation given below

$$p(x|\lambda) = \sum_{i=1}^{M} w_i \; g(x|\mu_i, \sum_i), \quad (1)$$

where x represents a D-dimensional continuous-valued data vector, w_i are the mixture of weights ,where $(i = 1, .., M)$ and $g(x|\mu_i, \sum_i)$, are the component gaussian densities, where $(i = 1, \dots, M)$.

There are some weaknesses in the HMM systems which affects the performance of the GMM-HMM as well, like conditional independence and HMMs trained with maximum likelihood criterion, which lacks in discriminative power [9].

2.2. Deep Neural Networks

Within the last few years, Deep Neural Network-Hidden Markov Model (DNN-HMM) based ASR systems becoming more popular due the advancement in techniques; DNNs can use HMM state alignment as outputs in place of hand labelled phones, can use pre-training to improve training accuracy of models using many hidden layers and are good at handling multiple frames of input coefficients [9,19-23]. In the DNN-HMM ASR architecture, the word transition weights are modelled using HMM, while the acoustic weights are modelled using DNN approach.

A Deep Neural Network (DNN) is a feed-forward, Artificial Neural Network (ANN); consists of multiple hidden layers between its inputs and its outputs. Mathematically, equation 2, represents, each hidden unit, j, that uses the logistic function to map the comprehensive input from the below layer,x_j, to the scalar state, y_j, which it sends to the above layer.

$$y_j = logistic(x_j) = \frac{1}{1 + e^{-x_j}}, x_j = b_j + \sum_i y_i w_{ij}, \quad (2)$$

Where b_j is the bias of unit and w_{ij} represents the weight on a connection to unit j from unit i in the layer below it [9]. A class probability, p_j in case of multiclass classification, can be defined as

$$p_j = \frac{\exp(x_j)}{\sum_k \exp(x_k)}, \quad (3)$$

Furthermore, there are certain components, which will affect the design of DNN acoustic models positively for Large Vocabulary Continuous Speech Recognition (LVSCR). Some of them are; varying the neural network architecture, using appropriate optimization algorithm, acoustic model training loss function, by handing the overfitting problem, the effect of DNN size on task performance, DNN training and choice of activation function also revolutionized the performance of ASR systems using deep learning approaches [26-27].

Recently, TDNNs gained popularity because of two major advantages in comparison to other existing GMM-HMM acoustic models; reduced training time as it uses sub-sampling and due to improved ASR systems; most recently due to the state-of-the-art TDNN-F [25]. Although the Recurrent Neural Networks (RNNs) are also doing well in the long temporal contexts, their training time is more than TDNNs, because RNNs uses the learning algorithm which takes the information in a sequential manner [30].

Furthermore, TDNN-F, uses low-rank factorized layers, structure wise it is very similar as the TDNN, also uses the Singular Value Decomposition (SVD) to compress the layers; which is the most popular method to minimize the number of parameters for pre-trained NNs [25].

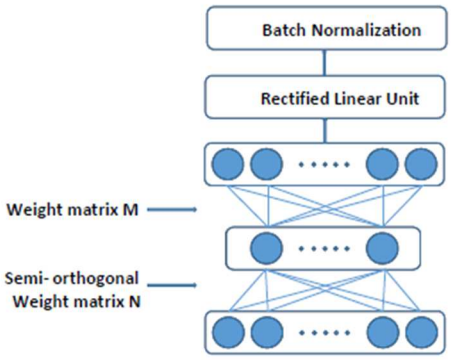

Figure 2: *Factorized layer of TDNN-F [25].*

3. Weighted finite state transducers

From the last few years, Kaldi become popular among ASR researchers and currently the state-of-the-art ASR toolkit [24]. Kaldi uses Weighted Finite State Transducers (WFSTs) to merge the information taken from the acoustic model and language model [24]. Table 1 shows, how WFSTs maps symbols or objects of an input alphabet to an object in the output alphabet

Table 1: *Components representing WFSTs*

Transducer	Input sequence	Output sequence
(G) word-level grammar	words	words
(L) pronunciation lexicon	phones	words
(C) context-dependency	CD phones	phones
(H) HMM	HMM states	CD phones

Table 1 represents, that composing L and G results in a transducer L $\circ$ G, which further maps a phone sequence into a word sequence. Finally, H $\circ$ C $\circ$ L $\circ$ G, outcomes in a transducer; which maps from the HMM states to a word sequence in a systematic manner.

4. Experimental setup

4.1. Database profile

There are two types of the dataset used for training and testing. Pure Malay speech corpus (MASS) [28]; used for training and Malay-English code-switched corpus used for the testing purpose, both with a sampling rate of 16 kHz. The entire datasets involve about 169 hours of speech signals, in which approximately 140 hours is the Malay speech and around 29 hours is the Malay-English speech. The lexicon contains 615125 words and pronunciations. The words are modelled using Malay phonemes. The pronunciation of English words used the closest Malay phonemes to the English phonemes in term of perception. An approximate number of phonemes of Malay and English language is mentioned in Table 2, as it varies in terms of dialect, especially in case of English.

The following is a Malay-English CS example taken from the Malay-English CS corpus;

Original: *selain itu normala juga sempat membisik dia punya rancangan untuk menerbitkan biografi mengenai dirinya sepanjang bergelar* public figure.

Translation: besides, normala also had the chance to describe her program to publish a biography about her as a public figure.

Training dataset consists of 200 speakers; out of which 82 are males and 118 are females while testing dataset consists of total 47 speakers; out of which 27 are males and 20 are females.

Additionally, data preparation is one of the important initial steps to use Kaldi for an ASR task; which is further divided into acoustic data and language data. It is mandatory to create "text", "spk2utt", "utt2spk" and "wav.scp" files, to use meta-data as acoustic data. Mainly "lexicon.txt", "nonsilence_phones.txt" and "silence_phones.txt" meta-data files are used for preparing language data. There are also several files need to be created to test the datasets using the Kaldi ASR toolkit.

Table 2: *Approximate number of Phonemes in Malay and English*

Lang.	Cons.	Mono.	Dip.	Tri.	Tot.
Malay	27	6	3	-	36
English	24	10	6-7	1-3	44

Table 3: *Training and testing corpus details*

Purpose	Lines	Persons	Size(GB)
Training	58421	200	16.0
Testing	12370	47	3.2

Following Table 4 shows the technical specifications of the speech recordings.

Table 4: *Technical details of the speech recordings*

Technical details
Channels : 1
Sampling Rate : 16 kHz
Precision : 16-bit
Bit rate : 256 kbit/s
Sample Encoding : Integer PCM

4.2. GMM-HMM training

This is the first step to start training, as we can do GMM-HMM training at utterance level transcriptions, but DNN training required labeled frames (phoneme-to-audio alignments) which were generated by a GMM-HMM system. The quality of GMM-HMM training does not depend on previous training, but DNN will be greatly affected by the quality of the GMM-HMM previously trained. An inferior quality GMM-HMM will give bad alignments, which will further affect DNN training badly Additionally, Mel-Frequency Cepstral Coefficient (MFCC) method used for feature extraction purpose, as it performs better than its counterparts [29].

4.2.1. Monophone training

Training monophone models is the very first and quick step of the training process. As monophone refer to a single phone and context independent in nature, it relies mainly on acoustic data and lexicon data.; mainly used bootstrap training for later models.

4.2.2. Triphone training

In triphones; each phone has a unique model for each left and right context. Initial triphone model (tri1) also trained in the same way as phone-based GMM-HMM model using the train_deltas script. Second triphone model, tri2b used Linear Discriminant Analysis + Maximum Likelihood Linear Transform (LDA+MLLT), because this combination can minimize the WERs, it can also boost up the decoding speed marginally, as it decreases the size of the acoustic model also. Other triphone models, tri3b used (LDA+MLLT+SAT), ti4b used (LDA+MLLT+SAT+SAT) and tri5b is larger SAT model.

4.3. TDNN training

After GMM-HMM training, i-vector feature extraction was performed, here i-vector training script extracts the i-vector on the speed-perturbed training data after combining short segments. Having a sizable number of speakers is helpful for generalization, and it also works fine in handing per-utterance decoding as well. Also, it will further improve the performance of TDNNs; as it contains the compatible information regarding a speaker's identity in a low dimensional fixed-length representation; broadly used in speaker adaptation of ASR. The DNN model we used here is the TDNN-F, it uses semi-orthogonal constraint; low-rank factorized layer; with lattice free MMI, it also contains total 11 hidden layers, with linear bottlenecks of 256, hidden layer dimensions varied from 1024 to 1536 and final layer is factorized [25]. Training run for 6 epochs, with an initial effective rate of 0.001 for optimization. A GPU; GeForce GTX 1070; contains 1920 CUDA cores also used to speed up the computations.

5. Results

In Table 5, triphone based GMM-HMM systems outperformed the monophone model. Among all triphone models, tri5b is the best, as it outperformed the monophone model with a relative percentage reduction of 29.83% in terms of WER.

Table 5: *WER of different acoustic models*

Acoustic Model	Description	WERs (%)
mono	Monophone	40.96
tri1	GMM-HMM	30.85
tri2b	LDA+MLLT	29.04
tri3b	LDA+MLLT+SAT	30.65
tri4b	LDA+MLLT+SAT+SAT	29.44
tri5b	Larger SAT model	28.74
TDNN-F	Factored form of TDNN	**20.67**

Although, the second best triphone model is tri2b, as it gives, 29.10% relative improvement, as compared to monophone model and stands quite close to tri5b.

Finally, the best WER; of our overall acoustic models was achieved by the TDNN-F; that is 20.67%, which is a 28.07%

relative percentage reduction in the WERs in comparison to tri5b; best triphone model. There is a huge 49.53 % relative improvement, by comparing TDNN-F with the monophone model.

6. Conclusions

In this paper, we have presented a new baseline for Malay-English code-switched corpus. In general, code-switched speech recognition is a difficult task, as the speaker can switch the language within the conversation in an unpredictable manner. Also, the best WER was achieved by the TDNN-F, with a 28.07% relative improvement with respect to the best triphone model, tri5b GMM-HMM system developed here, which shows that the factored form of TDNNs, along with a semi-orthogonal constraint is advantageous and can improve the accuracy for Malay-English code-switched corpus.

Methods and techniques are improving in a fast manner, especially in case of DNNs, so appealing future direction, mainly for low-resourced language will be to take advantage of the well resourced related language to develop better ASR systems. We will continue to improve the ASR performance, as the newly developed baseline, provides a good reference for future experiments.

7. Acknowledgements

This work is supported by Fundamental Research Grant (FRGS) 203.PKOMP.6711536 from Ministry of Higher Education Malaysia.

8. References

[1] R.P. Lippmann, "Speech Recognition by machines and humans," Speech Commun., vol. 22, no. 1, pp. 1-15, 1997.

[2] S. Young, "Talking to machines," Ingenia, vol. 54, pp. 40-46, 2013.

[3] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Google search by voice: A case study," in Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, chapter 4, pp. 61-90, Springer, 2010.

[4] T. Gulzar, A. Singh, D. K. Rajoriya and N. Farooq, "A systematic analysis of automatic speech recognition: An overview", International Journal of Current Engineering and Technology (IJCET), vol. 4, no. 3, pp. 1664-1675, 2014.

[5] K. Soky, V. Chea, and S. Sam, "Khmer automatic speech recognition system based on DNN models," in First Regional Conf. on Optical character recognition and Natural language processing for ASEAN Languages (ONA), Phnom Penh, Cambodia, 2017.

[6] L. Rabiner, and R. Schafer, "Introduction to digital speech processing," Foundations and Trends in Signal Processing, vol. 1, pp. 1-194, 2007.

[7] M. J. F. Gales, and S. Young, "The Application of Hidden Markov Models in Speech Recognition," Foundations and Trends in Signal Processing, vol. 1, no. 3, pp. 195-304, 2007.

[8] Y. Bengio, "Deep learning of representations: Looking forward", Proceedings of the First International Conference on Statistical Language and Speech Processing, pp. 1-37, 2013.

[9] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitley, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," Signal Processing Magazine, IEEE, vol. 29, no.6, pp. 82-97, 2012.

[10] L. Rabiner, and B.-H. Juang, Fundamentals of speech recognition, 1st ed. Prentice Hall, New Jersey, USA, 1993.

[11] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," Speech Communication, vol. 56, pp. 85-100, 2014.

[12] D. Fohr, O. Mella, and I. Illina, "New paradigm in speech recognition: deep neural networks," in IEEE Intl. Conf. on Information Systems and Economic Intelligence, 2017.

[13] R. Sahraeian, "Acoustic modeling of under-resourced languages," Ph.D. dissertation, Dept. of Electrical Eng., KU, LEUVEN, Belgium, 2017.

[14] Y.-L. Yeong, and T.-P. Tan, "Language identification of Malay-English words using syllable structure information," in Proc. Workshop Spoken Languages Technologies for Under-resourced Languages (SLTU), Penang, Malaysia, 2010.

[15] P. Auer, Code-switching in conversation: Language, interaction and identity, London, Routledge, 1998.

[16] P. Muysken, Bilingual Speech: A Typology of Code-mixing. Cambridge University Press, 2000.

[17] S. Thomason, Language Contact: An Introduction. Edinburgh University Press, 2000.

[18] E. Yilmaz, H. V. D. Heuvel, and D. V. Leeuwen, "Investigating bilingual deep neural networks for automatic speech recognition of code-switching Frisian speech," in Proc. Workshop on Spoken Language Technology for Under-resourced Languages (SLTU), pp. 159-166, 2016.

[19] M. Bhargava, and R. Rose, "Architectures for deep neural network based acoustic models defined over windowed speech waveforms," in Proc. of Interspeech, pp. 1-5, 2015.

[20] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Netw., vol. 61, pp. 85-117, 2015.

[21] L. Deng, "Deep learning: from speech recognition to language and multimodal processing," APSIPA Trans. on Signal Inform. Process., vol. 5, pp. 1-15, 2016.

[22] J. Pan, C. Liu, Z. Wang, Y. Hu, and H. Jiang, "Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling," in Proc. Int. Symp. Chinese Spoken Lang. Process. (ISCSLP), 2012.

[23] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in Proc. of the Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 215-219, 2014.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannenmann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely "The Kaldi speech recognition toolkit," in IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, 2011.

[25] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-Orthogonal low-rank matrix factorization for deep neural networks" in Interspeech, 2018 (submitted).

[26] H. K. Vydana, and A. K. Vuppala, "Investigate study of various activation functions for speech recognition" Twenty third National Conf. on Comm. (NCC), pp. 1-5, 2017.

[27] A. Mass, P. Qi, Z. Xie, A. Hannun, C.T. Lengerich, D. Jurafsky, and A.Y. Ng, "Building DNN acoustic models for large vocabulary speech recognition," Intl. J. Computer Speech and Lang. , vol. 41, pp. 195-213, 2017.

[28] T.-P. Tan, X. Xiao, E.S. Chng, E.K. Tang, and HZ. Li, "MASS: A Malay Language LVCSR Corpus Resource," Cocosda'09, Beijing, China, pp. 10-13, 2009.

[29] T. Gulzar, A. Singh and S. Sharma, "Comparative analysis of LPCC, MFCC and BFCC for the recognition of Hindi words using Artificial Neural Networks ", International Journal of Computer Applications (IJCA), vol. 101, no. 12, pp. 22-27, 2014.

[30] J. Leinonen, P. Smit, S. Virpioja, and M. Kurimo, "New baseline in automatic speech recognition for Northern Sámi," in Proc. Of the 4th Intl. Workshop for Computational Linguistics for Uralic Languages (IWCLUL), pp. 89-99, Helsinki, Finland, 2018.