Automatic Speech Recognition: Low Resource and Noise Robustness

Ankur T. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

ankur_patil@daiict.an.in

Abstract

Recently, developing Automatic Speech Recognition (ASR) systems for Low Resource (LR) languages, and distant multi-microphone conversational speech recognition, are active research area. The research in ASR is significantly advanced using deep learning approaches producing state-of-the-art results compared to the conventional approaches. However, it is still challenging to use such approaches for LR languages since it requires a huge amount of training data. Also, distant multi-microphone ASR remains a challenging goal in everyday environments involving multiple background sources and reverberation. Recently, deep learning based end-to-end ASR system developement approach is emerging along with RNNLM and robust feature representation improvement in the performance of the ASR. My research work is contributing in both of the above mentioned aspects of ASR as our lab participated in the Low Resource Speech Recognition Challenge for Indian Languages, and CHiME-5 Challenge in INTERSPEECH 2018, and bring out the significant improvement in the performance as compared to the baseline systems. Related developements in both aspects of ASR and our contribution is presented in this paper.

Index Terms: Low resource languages, ASR, deep learning, RNNLM.

1. Introduction

Speech is being the most natural way of communication, the research community is interested in Human Language Technologies (HLT) for human-machine interaction. It motivates the development of text-to-speech (TTS) and Automatic Speech Recognition (ASR) systems. ASR developement for low resource (LR) languages, and distant multi-microphone conversational speech recognition, are emerging research topics in HLT.

A language is considered as a LR when it lacks of unique writing system or stable orthography, linguistic expertise, electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, vocabulary lists, limited presence on the web, etc [1]. Only a small fraction of languages offers the resources required for the development of HLT [2]. Recently, deep learning aspects produce state-of-the-art results in many speech processing applications including ASR. The detailed survey on using deep learning for speech processing is presented in [3]. However, to train deep architectures, such as Deep Neural Networks (DNN), a large amount of training data is required. Hence, development of ASR using deep learning for LR languages is restricted due to the scarcity of data (in particular, audio, text or both). Still, use of high resourced and multilingual approaches have emerged to train DNN and consider LR language as a target language. To mitigate the limitation of developing ASR system for LR languages, we will either require innovative data collection/augmentation methodologies to increase the training data or the models for which information is shared amongst languages. There have been some efforts for the development of Indian language speech database for the Automatic Speech recognition (ASR) [4] and BABEL program [5]. Recently, ASR challenge for low resource Indian languages has been organized as a special session during the INTERSPEECH 2018. This challenge focuses on three Indian languages, namely, Gujarati, Telugu and Tamil. We participated in this challenge to develope ASR system for Gujarati language.

The developement in home automation and multimedia systems, has attracted the research in the distant multi-microphone conversational speech recognition which poses difficult reverberant and noisy conditions. The performance in this research topic has improved tremendously due to advances in speech processing, audio enhancement, and machine learning techniques. The CHiME challenges [6–8] and corpora have contributed to popularizing research on this topic, together with the DICIT [9], Sweet-Home [10], and DIRHA [11] corpora. The experimental results in CHiME-4 challenge shows the recent developement on this research topic. We perticipated in CHiME-5 challenge which has been organized as a satellite workshop during the INTERSPEECH 2018. Experimental setup and results of our ongoing work are discussed in next section.

2. Proposed System

2.1. Amplitude Modulation-based Features

The AM signals are extracted from the auditory filterbanks using the Energy Separation Algorithm (ESA) [12]. We have considered three type of filterbanks for our experiments. The two standard auditory filterbanks are gammatone and Gabor filterbank. The third one is obtained from the auditory filterbank learning using ConvRBM [13]. The block diagram for the AM feature extraction is shown in Figure 1. The short-time spectral features are obtained using framing with a Hamming window of squared envelopes followed by a logarithmic compression. The squaring and logarithm operation approximates the inner and outer hair cell nonlinearities, respectively in the cochlea [14].

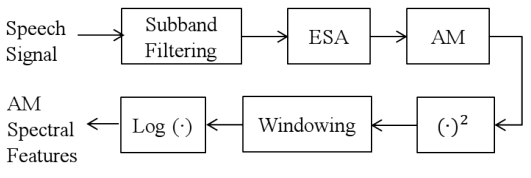

Figure 1: *Block diagram of AM spectral feature extraction.*

2.2. Recurrent Neural Networks for Language Modeling

Recurrent Neural Network Language Model (RNNLM) allows the information to persist by keeping loops in them [15]. We have used the Gated Recurrent Unit (GRU) as an activation function [16] and noise contrastive estimation (NCE) for the faster RNNLM training and testing [17]. The combination of RNNLM with n-gram LM is often done as shown in Figure 2 to preserve the essence of context and strong generalization. The LM probability using a linear interpolation of RNNLM with n-gram LM is given by [17]:

$$P(w_i|h_i) = \lambda P_{\mathbf{nG}}(w_i|h_i) + (1-\lambda)P_{\text{RNNLM}}(w_i|h_i), \quad (1)$$

where λ is a weight given to the n-gram LM $P_{\mathbf{nG}}(\cdot)$.

2.3. Deep Neural Networks for Acoustic Modeling

In this work, we used two DNN architectures which are Long Short-Term Memory (LSTM)-based RNN [18] and Time-Delay Neural Networks (TDNN). We also used TDNN-LSTM system which is recently proposed to get advantages of both TDNN and LSTM models [19]. In Figure 2, the TDNN/TDNN-LSTM block is shown which takes the labels from the Linear Discriminant Analysis (LDA)-Maximum Likelihood Linear Transform (MLLT) system. The decoding of the test data is performed using 3-gram LM followed by RNNLM rescoring.

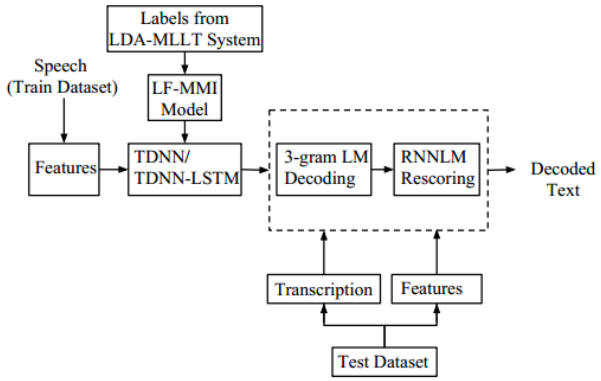

Figure 2: *Block diagram for the ASR system using neural networks.*

3. Experimental Setup

For the low resource speech recognition challenge for Indian languages, data is provided by SpeechOcean.com and Microsoft which is divided into a train and test sets [20]. The blind test set was released later as a part of the Challenge. We developed the ASR system for Gujarati language. The database of this language contains 22807, 3075, and 3419 utterances with 40, 5, and 5 hours of train, test and blind test data, respectively. While, CHiME-5 dataset is made up of the recording of twenty separate dinner parties taking place in real homes at different locations. Each party has been recorded with a set of six Microsoft Kinect devices. Each Kinect device has 4 sample-synchronised microphones and camera. Training data consists of recordings of 16 parties, while developement and evaluation data contains 2 parties each. This database contains 79980, 7440, and 11028 utterances with 40:33, 4:27, and 5:12 hours of train, developement and evaluation data, respectively. Participants are allowed to move naturally from one location to another and free to converse on any topics to make the conversational speech corpus.

For both databases, conventional GMM-HMM triphone system is built. We used the alignments obtained using the cepstral-based features for DNN training experiments with various filterbanks. We trained TDNN and TDNN-LSTM systems for DNN-based feature representation. For low resource challenge, Mel Frequency Cepstral Coefficients (MFCC) and AM-based cepstral features are extracted. The notations of AM cepstral features for three types of filterbanks are AM-GCC, AM-GTCC, and AM-ConvRBM-CC for Gabor, gammatone, and ConvRBM filterbanks, respectively. The ConvRBM-based filterbank (CBANK) is learned from the training database using the method proposed in [13]. The notations for AM spectral features for three types of filterbanks are denoted as AM-GTFB, AM-GFB, and AM-CBANK for gammatone, Gabor, and ConvRBM filterbanks, respectively.

The 3-gram LM is built using the SRILM toolkit [21] from the training corpus. The RNNLM is built with a training corpus in the Gujarati language using the faster-RNNLM toolkit [22]. The weight λ in the Eq. (1) is chosen to be 0.25, 0.5 and 0.75 for LM rescoring. All the ASR systems are trained in the Kaldi toolkit.

For CHiME-5 challege, they provided ene-to-end system over ESPnet, whereas we built it using Kaldi toolkit [23]. Speech enhancement is performed on training and developement data using weighted delay-and-sum beamforming technique.

4. Experimental Results

To explore the possible complementary information of various feature sets and classifiers, the system combination experiments (denoted as SC) are performed and reported in Table 1. The best performance is obtained with the SC-1 which includes combination of five ASR systems, (1) TDNN with FBANK, (2) TDNN with CBANK, (3) TDNN with AM-GFB, (4) TDNN-LSTM with AM-GTFB, and (5) TDNN-LSTM with CBANK. Using SC-1 combination strategy, there is a relative reduction of 4.3 % and 4.98 % over TDNN system trained with FBANK and decoded with RNNLM rescoring.

For CHiME-5 challenge, we built end-to-end system using Kaldi toolkit. Baseline end-to-end system gives 94.7% WER which is built using ESPnet. We are able to reduce the WER to 83.75% with ongoing experimentations.

Table 1: *Results of system various combination in % WER.*

System	Test	Blind Test
TDNN-FBANK (Baseline) [20]	19.76	28.99
TDNN-FBANK (Our baseline)	16.80	21.81
TDNN-FBANK with RNNLM	15.58	20.70
SC-5	**14.91**	**19.67**

5. Acknowledgement

I am very thankful to my Ph.D. supervisor Prof. Hemant A. Patil for his excellent guidance and motivation towards research publications. I also acknowledge NVIDIA for the hardware grant of TITAN X GPU to the Speech Research Lab, DA-IICT.

6. References

[1] S. Krauwer, "The basic language resource kit (BLARK) as the first milestone for the language resources roadmap," in *Proceed-*

ings of SPECOM, Moscow, Russia, 2003, pp. 8–15.

[2] 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.

[3] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, July 2017.

[4] G. Anumanchipalli, R. Chitturi, S. Joshi, R. Kumar, S. P. Singh, R. Sitaram, and S. P. Kishore, "Development of Indian language speech databases for large vocabulary speech recognition systems," in *International Conference on Speech and Computer (SPECOM), Patras, Greece*, 2005, pp. 591–594.

[5] IARPA, "The IARPA BABEL program," URL: https://www.iarpa.gov/index.php/research-programs/babel, {Last Accessed: 25 June 2018}.

[6] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The pascal chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[7] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second chimespeech separation and recognition challenge: Datasets, tasks and baselines," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 126–130.

[8] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 504–511.

[9] A. Brutti, L. Cristoforetti, W. Kellermann, L. Marquardt, and M. Omologo, "Woz acoustic data collection for interactive tv," *Language Resources and Evaluation*, vol. 44, no. 3, pp. 205–219, 2010.

[10] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "The sweet-home speech and multimodal corpus for home automation interaction," in *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, 2014, pp. 4499–4506.

[11] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The dirha-english corpus and related tasks for distant-speech recognition in domestic environments," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 275–282.

[12] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, 1993.

[13] H. B. Sailor and H. A. Patil, "Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 12, pp. 2341–2353, Dec. 2016.

[14] R. S. Schlauch, J. J. DiGiovanni, and D. T. Ries, "Basilar membrane nonlinearity and loudness," *The Journal of the Acoustical Society of America (JASA)*, vol. 103, no. 4, pp. 2010–2020, 1998.

[15] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH, Makuhari, Chiba, Japan*, 2010, pp. 1045–1048.

[16] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Deep Learning Workshop, NIPS 2014, Lake Tahoe, USA*, pp. 1–9, 2014.

[17] X. Chen, X. Liu, M. J. Gales, and P. C. Woodland, "Recurrent neural network language model training with noise contrastive estimation for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia*, 2015, pp. 5411–5415.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* 1^{st} *Edition*. The MIT Press, 2016.

[19] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and lstms," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2018.

[20] Microsoft, "INTERSPEECH 2018 special session: Low resource speech recognition challenge for Indian languages," URL: https://www.microsoft.com/en-us/research/event/interspeech-2018-special-session-low-resource-speech-recognition-challenge-indian-languages, 2018.

[21] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *International Conference on Spoken Language Processing (ICSLP),Colorado, USA*, 2002, pp. 901–904.

[22] Faster RNNLM, "Faster RNNLM (HS/NCE) toolkit," URL: https://github.com/yandex/faster-rnnlm, {Last Accessed: 18 March 2018}.

[23] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free mmi," pp. 1–5, 2018.