Interspeech 2018
2-6 September 2018, Hyderabad

DA-IICT/IIITV System for Low Resource Speech Recognition Challenge 2018

Hardik B. Sailor[1], *Maddala V. Siva Krishna*[2], *Diksha Chhabra*[2], *Ankur T. Patil*[1], *Madhu R. Kamble*[1] *and Hemant A. Patil*[1]

[1]Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India
[2]Indian Institute of Information Technology (IIIT), Vadodara, Gujarat, India

sailor_hardik@daiict.ac.in, 201551045@iiitvadodara.ac.in, 201451040@iiitvadodara.ac.in, ankur_patil@daiict.ac.in, madhu_kamble@daiict.ac.in, hemant_patil@daiict.ac.in

Abstract

This paper presents an Automatic Speech Recognition (ASR) system, in the Gujarati language, developed for Low Resource Speech Recognition Challenge for Indian Languages in INTERSPEECH 2018. For front-end, Amplitude Modulation (AM) features are extracted using the standard and data-driven auditory filterbanks. Recurrent Neural Network Language Models (RNNLM) are used for this task. There is a relative improvement of 36.18 % and 40.95 % in perplexity on the test and blind test sets, respectively, compared to 3-gram LM. Time-Delay Neural Network (TDNN) and TDNN-Long Short-Term Memory (LSTM) models are employed for acoustic modeling. The statistical significance of proposed approaches is justified using a bootstrap-based % Probability of Improvement (POI) measure. RNNLM rescoring with 3-gram LM gave an absolute reduction of 0.69-1.29 % in Word Error Rate (WER) for various feature sets. AM features extracted using the gammatone filterbank (AM-GTFB) performed well on the blind test set compared to the FBANK baseline (POI>70 %). The combination of ASR systems further increased the performance with an absolute reduction of 1.89 and 2.24 % in WER for test and blind test sets, respectively (100 % POI).

Index Terms: Gujarati language, RNNLM, amplitude modulation, TDNN, TDNN-LSTM.

1. Introduction

Speech and language technologies play a key role in a multilingual country, such as India. India has about 1652 native languages/dialects (even though there are only 22 official languages). Most of these official languages are still low-resourced. The Government of India is maintaining several resources in a web portal to increase the research and development of speech and language technologies [1]. There have been some efforts for the development of Indian language speech database for the Automatic Speech recognition (ASR) [2] and BABEL program [3]. Three low resource Indian languages, namely, Assamese, Bengali, and Tamil were included in the BABEL program [4]. A language is considered as low resource, when there is less or no availability of speech, text, phonetic dictionary, or transcribed data. To motivate the research in such languages, first of its kind 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. Gujarati is one of the official Indian languages which is still in the low resource category. Our earlier works in ASR for the Gujarati language include development of the phonetic engine for ASR [5] and ASR in the agricultural-domain [6] funded by MeitY, Govt. of India. In this paper, we have presented our Gujarati ASR system which is a part of the ASR Challenge.

Recently, there is a surge in the use of Recurrent Neural Network-based Language Model (RNNLM) for the ASR task. The detailed survey of RNNLM for LM is recently presented in [7]. Generally, RNNLM is used as a rescoring technique with n-gram LM [8]. There are various approaches for efficient training and testing using RNNLM, one of which we followed is proposed in [9]. We have also explored feature representations obtained from the multiband demodulation analysis (MDA) technique [10]. The AM and FM are two important physical aspects of communication sounds, such as speech signal. The AM-FM model of the speech describes dynamic changes in the envelope (AM) and carrier frequency (FM) [11]. For speech perception, the temporal envelope (AM) obtained from the subband filtering is essential as observed in psychoacoustics [12], [13] and the neurophysiological study in [14]. It was observed that AM is sufficient for speech recognition in clean conditions while FM does not provide any additional cues [15]. Since the challenge database has the clean conditions, we prefer only AM-based feature representation.

The objective of this paper is to show the effectiveness of RNNLM over 3-gram and AM-based spectral features for the ASR in Gujarati language. The RNNLM rescoring is performed with 3-gram LM. Two standard auditory filterbanks (Gabor and gammatone) and one data-driven auditory filterbank using Convolutional Restricted Boltzmann Machine (ConvRBM) [16] are used for the AM spectral feature extraction. The system combination of the proposed features along with RNNLM further improved the performance.

2. Amplitude Modulation-based Features

The AM signals are extracted from the auditory filterbanks using the Energy Separation Algorithm (ESA) [10]. The ESA algorithm estimates the instantaneous amplitude and frequency using the Teager Energy Operator (TEO) applied on the subband filtered signals [17]. The discrete version of the TEO (defined as $\Psi_D\{\cdot\}$) applied on the i^{th} subband $s_i[n]$ of the filterbank is defined as follows [10]:

$$\Psi_D\{s_i[n]\} := s_i^2[n] - s_i[n-1]s_i[n+1]. \qquad (1)$$

The discrete ESA-2 algorithm is used to extract the AM (i.e., envelope) $a_i[n]$ for the i^{th} subband as follows [18]:

$$a_i[n] \approx \frac{2\Psi_D\{s_i[n]\}}{\sqrt{\Psi_D\{s_i[n+1] - s_i[n-1]\}}}. \qquad (2)$$

We have considered three type of filterbanks for our experiments. The two standard auditory filterbanks are gammatone

10.21437/Interspeech.2018-1553

and Gabor filterbank. The third one is obtained from the auditory filterbank learning using ConvRBM [16]. It is shown earlier that data-driven ConvRBM filterbank contains complementary information compared to the standard filterbank [16], [19]. 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 [20].

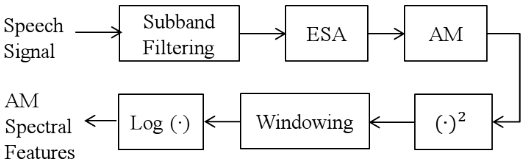

Figure 1: *Block diagram of AM spectral feature extraction.*

3. Neural Networks for Language and Acoustic Modeling

3.1. Recurrent Neural Networks for Language Modeling

Recurrent Neural Network Language Model (RNNLM) allows the information to persist by keeping loops in them [8]. It uses the previous information, $h_i = \{w_{i-1}, ..., w_1\}$ to predict the present word w_i. Its architecture consists of an input layer which is given a full history vector h_i containing the previous word w_{i-1} and vector v_{i-2} for remaining context. The hidden layer applies an activation function on the input and an output layer calculates the normalized RNNLM probabilities $P_{\text{RNNLM}}(w_i|w_{i-1}, v_{i-2})$ using a softmax layer. This process is repeated for calculating the probability of the next word w_{i+1} with the information being fed from the previous word. We have used the Gated Recurrent Unit (GRU) as an activation function in RNNLM [21]. RNNLMs are optimized using backpropagation through time (BPTT) algorithm with cross-entropy (CE) the objective function for training. In our study, we used the noise contrastive estimation (NCE) for the faster RNNLM training and testing [9]. 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 [9]:

$$P(w_i|h_i) = \lambda P_{\mathbf{nG}}(w_i|h_i) + (1 - \lambda)P_{\text{RNNLM}}(w_i|h_i), \quad (3)$$

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

3.2. Deep Neural Networks for Acoustic Modeling

In this paper, we consider to use DNN to model the temporal dynamics in the speech signal. Two such architectures are Long Short-Term Memory (LSTM)-based RNN and Time-Delay Neural Networks (TDNN). To model the sequential data, such as time series, speech, etc., RNN is the first choice. The most effective and popular sequence models are used in the practical applications called as gated RNNs which include the LSTM [22]. The LSTM model is based on introducing self-loops to produce the paths, where the gradient can flow for a longer duration. Using the gate controlled by the hidden unit, the time scale of integration can be changed dynamically [22]. Another DNN architecture which has been shown to be effective in modeling the long range temporal dependencies is the TDNN proposed in [23]. In TDNN, initial layers learn representations using narrow context whereas higher layers learn wider context [23]. TDNN is one of the best performing systems tested in the Kaldi toolkit for various ASR task. We also used TDNN-LSTM system which is recently proposed to get advantages of both TDNN and LSTM models [24]. For sequence-discriminative training of DNN acoustic models, we have used the Lattice-free Maximum Mutual Information (LF-MMI) in the HMM framework [25]. For better generalization, batch-normalization layers are added after TDNN layers. The L^2-regularization is also applied in the hidden and output softmax layer. 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.

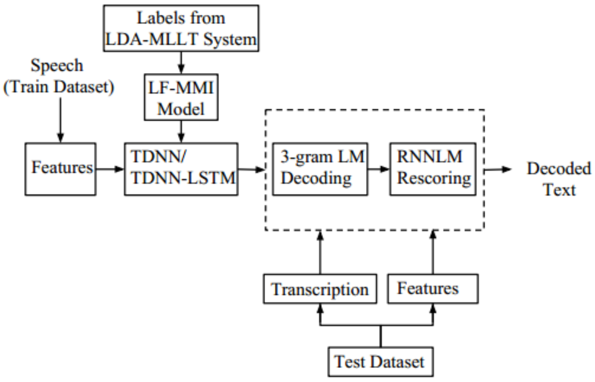

Figure 2: *Block diagram for the Gujarati ASR system using neural networks.*

4. Experimental Setup

4.1. Database

The ASR Challenge data is provided by SpeechOcean.com and Microsoft which is divided into a train and test sets [26]. The blind test set was released later as a part of the Challenge. The speech data is sampled at 16 kHz sampling frequency. Table 1 shows statistics of the Gujarati database. We have used the CMU Indic phoneset for the Gujarati language that consists of 54 phones [27]. The lexicon is provided by the challenge organizers.

Table 1: *Database for the Gujarati ASR system*

	Train	**Test**	**Blind Test**
Duration (hours)	40	5	5
No. of Utterances	22807	3075	3419

4.2. Feature Representations

For GMM-HMM training, Mel Frequency Cepstral Coefficients (MFCC) are extracted from the speech signals using a window length of 25 ms and shift of 10 ms. Delta and double-delta features are also appended resulting in 39-dimensional (D) features. The AM spectral features are extracted with 40 subband filters using the method shown in Figure 1. AM-based spectral representations are also converted into cepstral features to train GMM-HMM systems. 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. For DNN training, Mel filterbank (FBANK) and all the AM spectral features are used. The delta and double-delta features are appended resulting in 120-D features. The ConvRBM-based filterbank (CBANK) is learned from the training database using the method we proposed in [16]. Additionally, the annealed dropout is applied as done in [19] along with Leaky Noisy Rectifier Linear Units (LNReLU). 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.

4.3. GMM-HMM System Building

The GMM-HMM triphone system is built by varying the number of Gaussians and senones using 39-D cepstral features. The LDA-MLLT is applied to reduce the dimension and decorrelate the context-based cepstral features. The 3-gram LM is built using the SRILM toolkit [28] from the training corpus. The alignments obtained from the LDA-MLLT system are used in the hybrid DNN-HMM training. We used the alignments obtained using the cepstral-based features for DNN training experiments with various filterbanks.

4.4. Training of RNNLM and DNN

The RNNLM is built with a training corpus in the Gujarati language using the faster-RNNLM toolkit [29]. We have used 20 noise samples in the NCE training for the RNNLM. The number of hidden neurons and layers are selected based on its performance for the ASR task. The weight λ in the Eq. (3) is chosen to be 0.25, 0.5 and 0.75 for LM rescoring. All the ASR systems are trained in the Kaldi toolkit [30]. We trained TDNN with 1024 hidden neurons, 8 hidden layers and [-16,10] network context. The TDNN-LSTM system has three LSTM layers with 1024 cells, 256 recurrent/non-recurrent projection dimension and 9 TDNN layers of 1024 neurons. The L^2-regularization of 0.01 is applied in the hidden layers of both TDNN and TDNN-LSTM systems. For softmax output layer, L^2-regularization of 0.0025 and 0.004 is used for TDNN and TDNN-LSTM systems, respectively. The system combination is performed using the Minimum Bayes Risk (MBR) technique [31] with uniform weights to all the systems under consideration (i.e., hypothesis-level combination).

4.5. Evaluation Measures

The performance of an LM is reported in terms of the probabilistic measure called as perplexity (PPL) [32]. Lower the perplexity (i.e., higher the probability), better the language model performance and vice versa. For PPL computation, both the 3-gram and RNNLM are built from the training corpus and hence do not incorporate any prior knowledge of test set utterances or its vocabulary. We have also computed an increment in PPL to measure the LM performance when the difficulty-level is increased in the blind test set as compared to the test set. The ASR system performance is evaluated using the % Word Error Rate (WER). The statistical significance of one ASR system performing better than the other is assessed using % Probability of Improvement (POI) measure calculated using the bootstrap estimation of WER [33].

5. Experimental Results

5.1. Results of LM Evaluation

The experimental results of the performance of 3-gram and RNNLM are shown in Table 2. The PPL of the RNNLM is reported with 700 hidden neurons and two hidden layers (based on tuning the number of neurons and layers). The PPL of test and blind test sets are significantly lower for RNNLM. There is a relative improvement of 36.18 % on the test set and 40.95 % on the blind test set in the PPL using RNNLM compared to 3-gram. We have also shown the increased PPL from test to blind tests for both LMs. It is interesting to note that the PPL increment of RNNLM is 9.29 that is very low compared to 21.22 in 3-gram LM. Hence, RNNLM performs significantly better than 3-gram LM for our system.

Table 2: *Comparison of LMs using perplexity (PPL) as an evaluation metric*

Language Model	Test	Blind Test	Increased PPL
3-gram	68.02	89.24	21.22
RNNLM (700 × 2)	**43.41**	**52.70**	**9.29**

5.2. Results of GMM-HMM Experiments

The experimental results of the GMM-HMM experiments on the test set are shown in Table 3. Better results are obtained using 2800 senones and 22 Gaussians in both the triphone and LDA-MLLT systems for all the feature sets. Our MFCC baseline using the LDA-MLLT system has significantly lower % WER than the challenge baseline (2.56 % absolute reduction). RNNLM rescoring reduce the % WER for all the cepstral features compared to 3-gram LM. The AM-GTCC performed well compared to all the features with 20.14 % WER.

Table 3: *Results of GMM-HMM experiments on the test set using various feature sets in % WER*

Feature Set	3-gram	RNNLM
MFCC (Challenge Baseline) [26]	23.78	-
MFCC (Our Baseline)	21.22	20.35
AM-ConvRBM-CC	22.02	21.19
AM-GCC	**21.02**	20.45
AM-GTCC	21.03	**20.14**

5.3. Results of DNN-HMM Experiments

The experimental results of the test set using the DNN-HMM systems are reported in Table 4 in terms of % WER. In the case of 3-gram LM and TDNN system, AM-GFB performed better than FBANK feature set (62.01 % POI). Using TDNN-LSTM system, the AM-GTFB performed better than FBANK for both the 3-gram and RNNLM with 80.48 % and 74.65 % POI, respectively. After RNNLM rescoring, the performance of all the features are increased with an absolute reduction in a range of 1.05-1.58 % in WER for TDNN and 1-1.09 % for TDNN-LSTM system. The TDNN-LSTM system gave the improvement over TDNN for AM-GFB and AM-GTFB feature sets when used with RNNLM. However, FBANK with TDNN and RNNLM rescoring performed better on the test set compared to all the AM spectral features.

Table 4: *% WER using various features for the test set*

Feature Set	TDNN		TDNN-LSTM	
	3-gram	RNNLM	3-gram	RNNLM
FBANK	16.80	**15.58**	16.70	15.68
AM-CBANK	17.14	15.86	17.04	15.97
AM-GFB	**16.77**	15.72	16.75	15.66
AM-GTFB	16.82	15.66	**16.61**	15.61

The experimental results of the blind test set using the DNN-HMM systems are reported in Table 5 in terms of % WER. The AM-GTFB gave lower % WER compared to the FBANK when used in TDNN system (77.43 % POI). Using TDNN-LSTM system, the AM-GTFB performed better than the FBANK with 99.74 % POI using 3-gram and 96.78 % POI using RNNLM. After RNNLM rescoring, the performance of all the features are increased with an absolute reduction of around 1 % in WER for both the DNN systems. The best performing system on the blind test set is the TDNN-LSTM system trained with the AM-GTFB feature set.

We have observed that AM-GTFB performed well for both the test sets and the results are statistically significant over FBANK features (POI>70 %). The GTFB is developed to mimic the human auditory filter shapes and frequency scale [34]. Hence, the AM features obtained from the GTFB performed better compared to the other feature sets. The auditory filterbank learning using ConvRBM shows that model learns 30 low-frequency subband filters representing frequency less than 5 kHz and only 10 subband filters to represent a frequency above 5 kHz (not shown here). This may be due to either less number of speakers available in the database so that ConvRBM is biased towards speaker-specific low pitch frequency (F_0) and its harmonics (i.e., $kF_0, k \in \mathbb{Z}^+$). Frequency range of phonetic sounds in the speech signals of the Gujarati language spans mostly in lower frequency regions. Hence, AM features from ConvRBM filterbank, i.e., AM-CBANK did not perform well even though it is obtained in a data-driven manner. However, later we observed that it captures significant complementary information in a system combination framework.

Table 5: *% WER using various feature set for blind test set*

Feature Set	TDNN		TDNN-LSTM	
	3-gram	RNNLM	3-gram	RNNLM
FBANK	21.81	20.70	22.00	20.77
AM-CBANK	22.22	21.07	22.39	21.49
AM-GFB	21.81	20.64	22.10	21.00
AM-GTFB	21.81	**20.61**	**21.70**	**20.57**

5.4. Results of System Combination Experiments

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 6. The comparison of our baseline with the challenge baseline is also shown here. Our FBANK baseline has significantly lower WER compared to the challenge baseline. We have tried various system combinations of feature sets and DNN systems out of which some of the combinations are shown in Table 6. We compared the system combination experiments with TDNN system trained with FBANK and decoded with RNNLM rescoring. First, we combined the TDNN systems (SC-1) trained with various feature sets (used in this study) that resulted in a relative improvement of 1.93 % for test (99.94 % POI) and 2.46 % for blind test (100 % POI). The combination of TDNN-LSTM system (SC-2) improves the performance on the test set only. The SC-3 combination includes two TDNN with FBANK and AM-GTFB, and one TDNN-LSTM with AM-GTFB. The SC-4 combination includes two TDNN with FBANK and AM-GFB, and two TDNN-LSTM with AM-GTFB and AM-CBANK. Both SC-3 and SC-4 combinations slightly improved the performance compared to SC-1 and SC-2. To get complementary information from ConvRBM-based filterbanks (CBANK), we have also used filterbank features directly (without AM processing) in rest of the combinations. The best performance is obtained with the SC-5 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-5 combination strategy, there is a relative reduction of 4.3 % and 4.98 % over TDNN system trained with FBANK and decoded with RNNLM rescoring (100 % POI).

Table 6: *Results of system various combination in % WER. Numbers in the round parenthesis indicates % POI calculated with reference to TDNN-FBANK with RNNLM rescoring.*

System	Test	Blind Test
TDNN-FBANK (Baseline) [26]	19.76	28.99
TDNN-FBANK (Our baseline)	16.80	21.81
TDNN-FBANK with RNNLM	15.58	20.70
SC-1	15.28 (99.94)	20.19 (100)
SC-2	15.25 (99.67)	20.28 (96.04)
SC-3	15.02 (100)	19.84 (100)
SC-4	14.98 (100)	19.82 (100)
SC-5	**14.91** (100)	**19.67** (100)

6. Summary and Conclusions

In this study, we have presented the development of the ASR system in the Gujarati language as a part of INTERSPEECH 2018 challenge on low resource speech recognition in Indian languages. The effectiveness of RNN-based language modeling over 3-gram LM is shown with perplexity as evaluation criteria. The AM-based spectral features obtained from standard filterbanks and representation learning-based approach are used along with the standard Mel filterbanks. The system combination of various feature sets used in TDNN/TDNN-LSTM networks shows that proposed approach gives lower % WER. The statistical significance of the proposed approach is also presented. Our future work includes further development of RNNLM. We would also like to explore other high-level auditory features and the acoustical analysis of Gujarati language.

7. Acknowledgements

Authors would like to thank the MeitY, Govt. of India, for two sponsored projects (1) TTS Phase-II, (2) ASR Phase-II and the authorities of DA-IICT. We acknowledge SpeechOcean.com and Microsoft for ASR Challenge data. We also acknowledge NVIDIA for the hardware grant of TITAN X GPU.

8. References

[1] TDIL, "Indian language technology proliferation and deployment centre," URL: http://tdil-dc.in/index.php?lang=en, {Last Accessed: 18 March, 2018}.

[2] 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.

[3] IARPA, "The IARPA BABEL program," URL: https://www.iarpa.gov/index.php/research-programs/babel, {Last Accessed: 18 March 2018}.

[4] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nussbaum-Thom, M. Picheny, Z. Tske, P. Golik, R. Schlter, H. Ney, M. J. F. Gales, K. M. Knill, A. Ragni, H. Wang, and P. Woodland, "Multilingual representations for low resource speech recognition and keyword search," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, Arizona*, Dec. 2015, pp. 259–266.

[5] K. D. Malde, B. B. Vachhani, M. C. Madhavi, N. H. Chhayani, and H. A. Patil, "Development of speech corpora in Gujarati and Marathi for phonetic transcription," in *International Conference Oriental COCOSDA held jointly with CASLRE*, Gurgaon, India, 2013, pp. 1–6.

[6] H. B. Sailor, H. A. Patil, and A. Rajpal, "Unsupervised filterbank learning for speech-based access system for agricultural commodity," in *IEEE International Conference on Advances in Pattern Recognition (ICAPR), Kolkata, India*, 2017, pp. 1–6.

[7] W. D. Mulder, S. Bethard, and M. F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Computer Speech & Language*, vol. 30, no. 1, pp. 61 – 98, 2015.

[8] 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.

[9] 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.

[10] 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.

[11] H. Luo, Y. Wang, D. Poeppel, and J. Z. Simon, "Concurrent encoding of frequency and amplitude modulation in human auditory cortex: MEG evidence," *Journal of Neurophysiology*, vol. 96, no. 5, pp. 2712–2723, 2006.

[12] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[13] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[14] O. Ghitza, A. L. Giraud, and D. Poeppel, "Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence," *Frontiers in Human Neuroscience*, vol. 6, p. 340, 2013.

[15] F. G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proceedings of the National Academy of Sciences*, vol. 102, no. 7, pp. 2293–2298, 2005.

[16] 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.

[17] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, New Mexico, USA, 1990, pp. 381–384.

[18] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* 1^{st} *Edition.* Pearson Education India, 2006.

[19] H. B. Sailor and H. A. Patil, "Auditory feature representation using convolutional restricted Boltzmann machine and Teager energy operator for speech recognition," *Journal of Acoustical Society of America Express Letters (JASA-EL)*, vol. 141, no. 6, pp. EL500–EL506, June. 2017.

[20] 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.

[21] 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.

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

[23] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH, Dresden Germany*, 2015, pp. 2440–2444.

[24] 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.

[25] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *INTERSPEECH, San Francisco, CA, USA*, 2016, pp. 2751–2755.

[26] 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.

[27] A. Parlikar, S. Sitaram, A. Wilkinson, and A. W. Black, "The festvox Indic frontend for grapheme-to-phoneme conversion," in *Workshop on Indian Language Data-Resources and Evaluation (WILDRE) under LREC 2018, Reykjavik, Iceland*, 2016, pp. 1–6.

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

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

[30] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Big Island, Hawaii, USA*, 2011, pp. 1–4.

[31] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802 – 828, 2011.

[32] F. Jelinek, R. Mercer, L. R Bahl, and J. K Baker, "Perplexity - a measure of the difficulty of speech recognition tasks," *Journal of the Acoustical Society of America (JASA)*, vol. 62, p. S63, 11 1977.

[33] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Montreal, Quebec, Canada*, 2004, pp. 409–411.

[34] S. Strahl and A. Mertins, "Analysis and design of gammatone signal models," *The Journal of the Acoustical Society of America (JASA)*, vol. 126, no. 5, pp. 2379–2389, 2009.