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

Signal Processing Cues to Improve Automatic Speech Recognition for Low Resource Indian Languages

Arun Baby, Karthik Pandia D S, Hema A Murthy

Indian Institute of Technology Madras
arunbaby@cse.iitm.ac.in

Abstract

Building accurate acoustic models for low resource languages is the focus of this paper. Acoustic models are likely to be accurate provided the phone boundaries are determined accurately. Conventional flat-start based Viterbi phone alignment (where only utterance level transcriptions are available) results in poor phone boundaries as the boundaries are not explicitly modeled in any statistical machine learning system. The focus of the effort in this paper is to explicitly model phrase boundaries using acoustic cues obtained using signal processing. A phrase is made up of a sequence of words, where each word is made up of a sequence of syllables. Syllable boundaries are detected using signal processing. The waveform corresponding to an utterance is spliced at phrase boundaries when it matches a syllable boundary. Gaussian mixture model - hidden Markov model (GMM-HMM) training is performed phrase by phrase, rather than utterance by utterance. Training using these short phrases yields better acoustic models. This alignment is then fed to a DNN to enable better discrimination between phones. During the training process, the syllable boundaries (obtained using signal processing) are restored in every iteration. A relative improvement is observed in WER over the baseline Indian languages, namely, Gujarati, Tamil, and Telugu.

Index Terms: automatic speech recognition, Indian languages, under resourced languages, signal processing cues, group delay, deep neural networks, long short-term memory

1. Introduction

Majority of Indian languages are digitally low resource [1]. Speech interfaces, especially in the vernacular, are enablers in such an environment. Automatic speech recognition (ASR) systems require transcribed speech from many speakers, pronunciation dictionaries, and massive amounts of text data to train statistical language models. In languages like English and Chinese, the amount of digital data available is large [2–4]. This is responsible for robust ASR systems for English and Chinese [5–7]. While building speech systems for Indian languages, the scarcity of the data is a primary concern. Attempts have been made to bootstrap training data for low resource languages [8]. Signal processing cues that are agnostic to the speaker is used in different speech systems to improve the performance [9–14].

A typical ASR system consists of an acoustic model and a language model. The language model is trained using the text data available. The acoustic model is trained using the input acoustic features. Robustness of the acoustic models depends on the accuracy of phone boundaries.

Phones are the most common subword unit for speech modeling. But in most of the cases, only sentence level transcription is available for training the models [2, 15, 16]. Obtaining accurate phone level alignment is a difficult task. Manual alignment is not only time-consuming but can also be inconsistent as it is difficult to perceive a phone in isolation. In [17], Siniscalchi et al. propose a common set of fundamental units that can be defined "universally" across all spoken languages.

Syllable, the fundamental unit of speech production can be used as an alternative to the phone. Syllables have typical spectral and temporal characteristics, and much longer in duration (about 150ms) and can be detected using signal processing. Syllables are also closely related to human speech perception and articulation [18]. Analysis of pronunciation variation at syllable level is observed to be more systematic [19]. Syllable is found to be a robust subword unit for Indian languages [20–22]. Syllable modeling results in the reduction of model parameters as context dependencies are less important for syllable models compared to that of tri-phone models [23, 24].

Syllable boundary detection from the acoustic waveform is comparatively easy since a syllable is characterized by an onset, nucleus, and coda as shown in Figure 1. Group delay (GD) based techniques have been shown to give robust syllable boundaries for Indian languages [9–12].

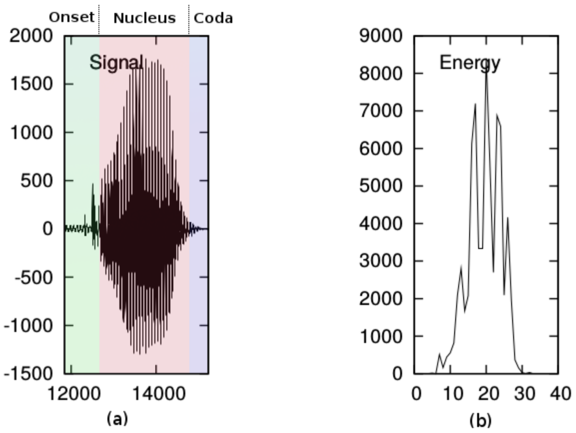

Figure 1: *Waveform and energy plots of a syllable. X-axis correspond to time in samples at a sampling rate of 16 kHz (fig a). The energy plot shows the corresponding number of frames in a syllable (fig b).*

In this work, first, energy and spectral flux are used to obtain reliable syllable boundaries. The initial mono phone training is restricted to phrases to create robust models. As the phone boundaries are not explicitly modeled in a deep neural network (DNN) system, the boundaries (obtained using spectral cues) are restored using an iterative correction approach.

The rest of the paper is organized as follows. Section 2 details the spectral cues used for detecting syllable boundaries. The proposed system is explained in Section 3. The experiments conducted are detailed in Section 4. The work is concluded in Section 5.

10.21437/SLTU.2018-6

2. Spectral cues

A syllable is of the form C*VC*, where C is a consonant and V is a vowel. A syllable consists of three parts - onset, nucleus, and coda. The onset and coda consist of consonants whereas nucleus is a vowel. A syllable consists of one or more phones. Owing to co-articulation in continuous speech, it is more difficult to distinguish phone transitions than syllable transitions [25]. It is easier to obtain syllable boundaries than phone boundaries. The region of vowels in syllables corresponds to more energy and duration than that of consonants. Boundaries of syllables correspond to low energy region. Short term energy (STE) can be used as an acoustic cue to obtain syllable boundaries. But STE cannot be used directly due to local fluctuations in energy (Figure 1). STE can be made to resemble magnitude spectrum of any real signal, and GD based processing can be applied to obtain the syllable boundaries. GD function can be applied to minimum phase signals only. Hence, the signal is made minimum phase by processing in the root cepstral domain (inverse DFT of the short-term energy function) [26, 27] and then GD function is applied on this minimum phase signal. In [12], the use of minimum phase group delay functions in finding syllable boundaries for speech recognition is proposed.

This GD based algorithm is agnostic to text transcription as the boundaries are obtained directly from the waveform independent of the transcription. The number of syllable boundaries given by the algorithm depends on the size of the Hanning window chosen in the cepstral domain [12], which in turn depends on a parameter called window scale factor (WSF). WSF is inversely proportional to the syllable rate of the utterance. Figure 2 shows the GD of the STE of a part of an utterance for various WSF values: $10, 3.4$, and 1 in the three panes below the waveform. From the figure, it can be observed that WSF of 3.4 gives good syllable boundaries.

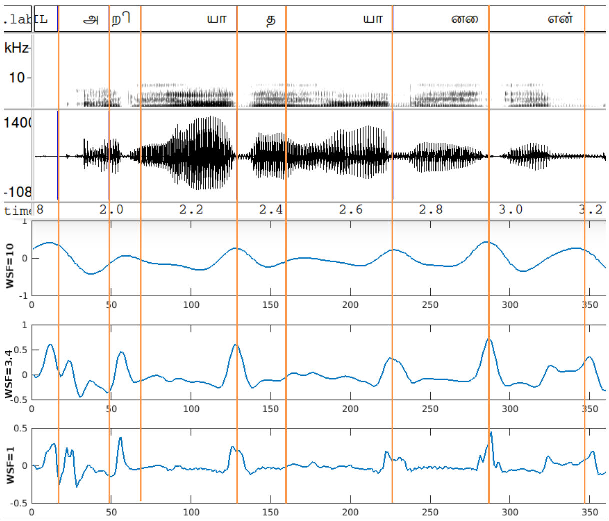

Figure 2: *GD boundaries for different WSF values*

Based on extensive experimentation, it is observed that sibilant fricatives and affricates have prominent energy in higher frequency bands, and nasals have prominent energy only in lower frequency bands. Spectral change as a function of time can be used to detect phone boundaries when the phone transition is accompanied by significant change in spectral characteristics [28, 29]. Spectral flux (SF), which is the Euclidean distance between the normalized power spectrum of a speech frame and normalized power spectrum of the previous frame, gives a measure of spectral change.

$$SF_n = (E_n - E_{n-1})^2$$

The peaks in the spectral flux correspond to phone boundaries. This property of spectral flux can be used for obtaining the phone boundaries of sibilant fricatives, and affricates [25]. Phone boundaries are characterized by energy changes in different bands of the spectrum [28]. Sub-band spectral flux (SBSF) is computed by dividing the normalized power spectrum into four bands uniformly, and finding the squared difference between the band energy of a frame with that of the previous frame as given by the equation:

$$SBSF_n = \sum_{n=1}^{4} (E_n[i] - E_{n-1}[i])^2$$

The boundaries obtained from GD processing and SBSF may not always yield correct boundaries. An observation with respect to Indian languages has led to the following set of rules that were developed for boundary marking between two syllables for building text to speech synthesis systems [25]:

Rule 1: The boundary between the syllables syllable1 and syllable2 is marked as correct (using STE) if the end phone of syllable1 is not a fricative or nasal, and the beginning phone of syllable2 is not a fricative, affricate, nasal or a semi-vowel.

Rule 2: The boundary between syllable1 and syllable2 is marked as correct (using SBSF) if the end phone of syllable1 or the begin phone of syllable2, but not both, is a fricative or an affricate.

3. Proposed system

The overview of the proposed system is shown in Figure 3. The system mainly has 5 modules. Each of the blocks is detailed below.

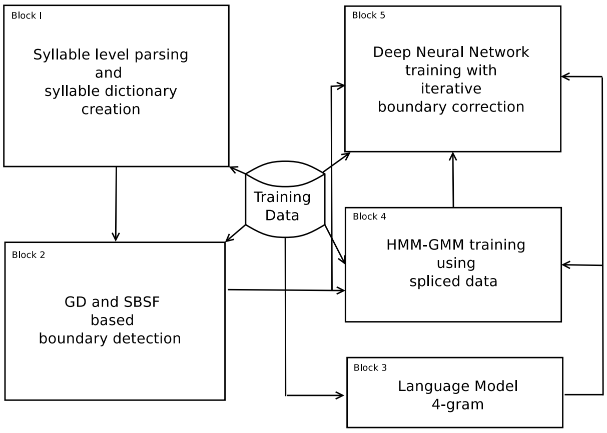
Block 1
Syllable level parsing and syllable dictionary creation
Block 5
Deep Neural Network training with iterative boundary correction
Training Data
Block 4
HMM-GMM training using spliced data
Block 2
GD and SBSF based boundary detection
Block 3
Language Model 4-gram

Figure 3: *Overview of the proposed system*

3.1. Syllable level parsing

The Block 1 in Figure 3 mainly deals with syllable level parsing and syllable dictionary creation. The unified parser which is a language-independent parser developed for Indian languages is used for this purpose [30]. The input text is split into words and parsed to obtain the corresponding syllable sequence. A syllable to phone dictionary is also created which will be later used in the group delay based boundary correction. This is shown in the Figure 4.

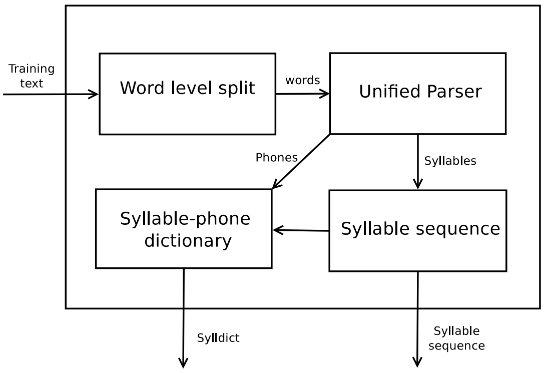

Figure 4: *Syllable level parsing*

3.2. GD and SBSF based boundary detection

Figure 5 shows the details of boundary detection with the help of spectral cues. In this module, syllable and word boundaries are obtained with the help of group delay processing [25]. The syllable sequence obtained from the previous module (Block 1) is used to train syllable models using flat-start initialization. MFCC features obtained using 25ms frame-size and 4ms frame-shift are used as the input features for training. HTK toolkit is mainly used in this module [31]. 14 iterations of embedded re-estimation are performed to get the final alignment.

Group delay based processing of STE and SBSF is performed using a fixed WSF. WSF is set as 6, which corresponds to the average syllable rate across Indian languages. The rules for corrections mentioned in Section 2 are used to find accurate syllable level boundaries. These boundaries are corrected in the alignment obtained. The silence (SIL) labels which are less than 20ms are removed from the alignment.

From the forced alignment obtained above, the syllable rate for each utterance is calculated. The average syllable rate per utterance is calculated after removing the silence part of the utterance.

With the newly obtained syllable rate, the WSF is adjusted for each utterance and the group delay based correction is repeated. Using the rules from Section 2 the appropriate syllable boundaries are corrected to obtain the syllable level boundaries. The syllables are combined later to obtain word boundaries.

3.3. Language modeling

The SRILM toolkit is used to train the language model [32]. 4-gram models are learned to build the language model.

3.4. GMM based training

The GMM based modeling is performed in this module which is shown in Figure 6. The utterances are spliced at reliable word boundaries (GD boundaries) to obtain sub-utterances. The spliced sub-utterances are then used to train the mono-phone models. Kaldi-toolkit is used [33]. Since the splicing is performed at reliable boundaries, phone models obtained using these sub-utterances will be better. Once the mono-phone models are trained the forced alignment is obtained using these models. These alignments are used further to train tri-phone models. Since the tri-phone models need context for better performance, the un-spliced training data is used. The TIMIT recipe of the Kaldi-toolkit is used for all tri-phone models. The final alignment obtained from tri3 models is fed as the input to the DNN systems.

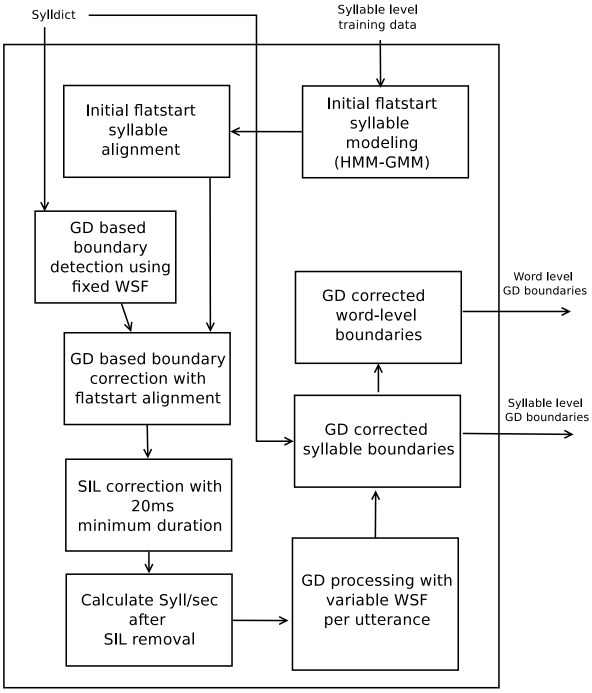
Sylldict
Syllable level training data
Initial flatstart syllable alignment
Initial flatstart syllable modeling (HMM-GMM)
GD based boundary detection using fixed WSF
GD based boundary correction with flatstart alignment
SIL correction with 20ms minimum duration
Calculate Syll/sec after SIL removal
GD corrected word-level boundaries
Word level GD boundaries
GD corrected syllable boundaries
Syllable level GD boundaries
GD processing with variable WSF per utterance

Figure 5: *Group delay based boundary detection*

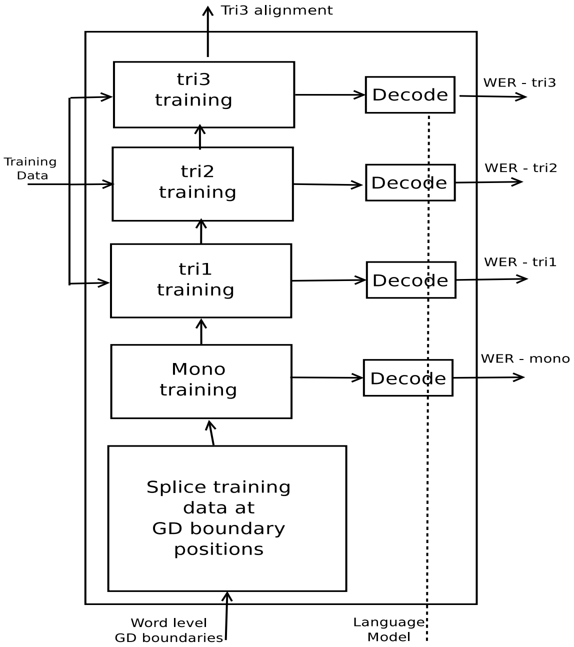

Figure 6: *GMM based training*

3.5. DNN based training

The iterative boundary correction using the GD corrected boundaries is performed here. The complete process is shown in Figure 7. The tri3 alignment from the previous block is used for the initial DNN training. The standard DNN configuration from the TIMIT recipe is used here for modeling. Once the DNN is trained, the alignment is obtained using these models. The alignment correction is performed using the GD syllable boundaries. Since the frame-shift used for input feature extraction is 25ms frame-size and 10ms frame-shift, there is a mismatch with the number of frames with the GD corrected boundaries. In order to overcome this, the GD boundaries are approximated to the closest 10ms boundary. These boundaries are then used to correct the alignments iteratively. These corrected alignments are again fed back to DNN to train again in an iterative fashion. The corrected boundaries are also fed as input to a long short-term memory (LSTM) training (TDNN+LSTM). The LSTM recipe is obtained from WSJ of Kaldi-toolkit.

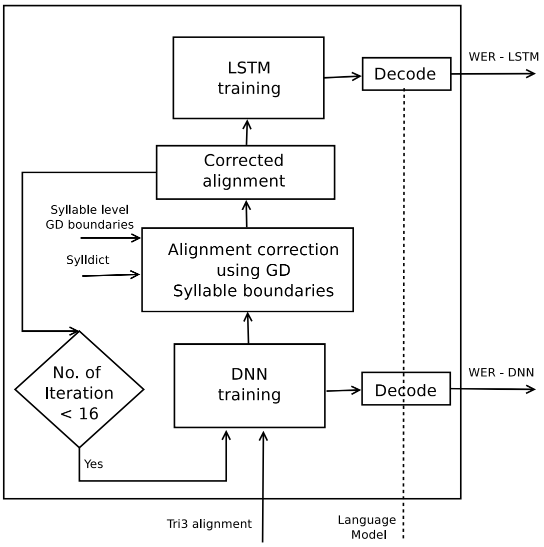

Figure 7: *DNN based training*

4. Experiments and Results

The dataset used for experimentation is released as part of “Low Resource Speech Recognition Challenge for Indian Languages” for INTERSPEECH 2018 by SpeechOcean.com and Microsoft. This contains data for 3 Indian languages: Gujarati, Tamil, and Telugu. The training data consists of 40 hours of read/conversational speech. The test data is of size 5 hours. The speaker information is unknown. Extensive experimentation is done for assessing the performance of different systems mentioned in Section 3.

Table 1 shows the number and percentage of the boundaries accurately detected using spectral cues. It shows that about 15−30% of the total boundaries are detected accurately using the spectral cues. The word level boundaries are used for training mono-phone models. The syllable level boundaries are used to correct the DNN alignments iteratively.

A baseline system is trained using the data provided. The performance of the proposed system is compared with this baseline system. The word error rate (WER) is used as the evaluation metric. Table 2 shows the WER for the different systems without iterative correction. It is observed that even though the initial WERs are bad for the proposed system, once it reaches the tri3 stage, a clear difference can be seen (marked in bold). But this difference vanishes as training progresses to the DNN and LSTM stages.

Table 1: *Boundary detection statistics*

	Tamil		Telugu		Gujarati	
Word level						
STE	30678	15.80%	19417	9.72%	43186	16.85%
SBSF	10479	5.40%	14394	7.20%	27694	10.80%
Total	194126		199798		256379	
Syllable level						
STE	98478	13.76%	74439	10.04%	98965	16.24%
SBSF	34858	4.87%	43694	5.90%	54942	9.01%
Total	715461		741129		609519	

Table 2: *WER for the proposed system without iterative correction*

Lang	Tamil		Telugu		Gujarati	
System	baseline	proposed	baseline	proposed	baseline	proposed
mono	50.18	54.46	48.08	49.51	33.33	35.29
tri1	32.36	37.35	32.56	37.37	23.34	26.77
tri2	29.09	33.35	29.76	34.45	21.54	24.97
tri3	27.86	**27.43**	29.32	**28.62**	21.25	**19.77**
DNN	22.43	22.09	23.98	23.71	18.49	17.79
LSTM	19.85	19.86	20.87	20.73	15.27	15.29

The performance of the DNN and LSTM systems are improved with the help of iterative correction of boundaries with the aid of signal processing cues. Iterative correction is performed 16 times. Table 3 shows the WER for top three systems using the iterative corrected DNN and LSTM systems. A clear improvement in WER is observed compared to the baseline.

Table 3: *Top 3 WERs for the proposed system with iterative correction*

Tamil		Telugu		Gujarati	
DNN	LSTM	DNN	LSTM	DNN	LSTM
22.09	19.47	23.62	20.32	17.60	15.04
22.13	19.56	23.67	20.38	17.71	15.05
22.15	19.62	23.68	20.46	17.72	15.08

5. Conclusion

Building robust ASR systems for Indian languages is a difficult task owing to the low resource nature of these languages. HMM-DNN/CTC systems which are state-of-art for ASR require huge amounts of data. We show that acoustic-phonetic cues obtained by processing the raw signal can be used to iteratively improve the performance of HMM-DNN systems.

6. Acknowledgements

Authors would like to thank SpeechOcean.com and Microsoft for collecting and providing the data as part of “Low Resource Speech Recognition Challenge for Indian Languages”, INTERSPEECH 2018.

7. References

[1] Wikipedia, "Languages of India - wikipedia, the free encyclopedia," 2018, [Online; accessed 22-March-2018]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Languages_of_India&oldid=831676831

[2] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[3] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[4] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8126–8130.

[5] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[7] A. Yuan, N. Ryant, N. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models," in *INTERSPEECH, ISCA (2013)*, 2013, pp. 2306–2310.

[8] X. Cui, J. Xue, X. Chen, P. A. Olsen, P. L. Dognin, U. V. Chaudhari, J. R. Hershey, and B. Zhou, "Hidden markov acoustic modeling with bootstrap and restructuring for low-resourced languages," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, pp. 2252–2264, 2012.

[9] H. A. Murthy and B. Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana*, vol. 36, no. 5, pp. 745–782, 2011.

[10] S. A. Shanmugam and H. Murthy, "A hybrid approach to segmentation of speech using group delay processing and hmm based embedded reestimation." in *INTERSPEECH*, 2014, pp. 1648–1652.

[11] T. Nagarajan and H. A. Murthy, "Subband-based group delay segmentation of spontaneous speech into syllable-like units," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 17, pp. 1–12, 2004.

[12] V. K. Prasad, T. Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communication*, vol. 42, no. 3, pp. 429–446, 2004.

[13] A. Baby, J. J. Prakash, R. Vignesh, and H. A. Murthy, "Deep learning techniques in tandem with signal processing cues for phonetic segmentation for text to speech synthesis in indian languages," in *Proc. Interspeech 2017*, 2017, pp. 3817–3821. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-666

[14] A. S. J. J. P. R. K. K. R. V. S. Anusha Prakash, Arun Baby and H. A. Murthy, "Blizzard challenge 2015: Submission by donlab, IIT Madras." [Online]. Available: http://www.festvox.org/blizzard/bc2015/DONLab_IITM_bc2015.pdf

[15] A. Baby, A. L. Thomas, N. L. Nishanthi, and T. Consortium, "Resources for indian languages," in *CBBLR – Community-Based Building of Language Resources*. Brno, Czech Republic: Tribun EU, Sep 2016, pp. 37–43.

[16] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[17] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[18] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on speech and audio processing*, vol. 9, no. 4, pp. 358–366, 2001.

[19] S. Greenberg, "Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2-4, pp. 159–176, 1999.

[20] H. A. Patil, T. B. Patel, N. J. Shah, H. B. Sailor, R. Krishnan, G. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra *et al.*, "A syllable-based framework for unit selection synthesis in 13 indian languages," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*. IEEE, 2013, pp. 1–8.

[21] A. Pradhan, A. Prakash, S. A. Shanmugam, G. Kasthuri, R. Krishnan, and H. A. Murthy, "Building speech synthesis systems for indian languages," in *Communications (NCC), 2015 Twenty First National Conference on*. IEEE, 2015, pp. 1–6.

[22] A. Lakshmi and H. A. Murthy, "A syllable based continuous speech recognizer for tamil," in *Ninth International Conference on Spoken Language Processing*, 2006, pp. 1878–1881.

[23] M. Y. Tachbelie, S. T. Abate, L. Besacier, and S. Rossato, "Syllable-based and hybrid acoustic models for amharic speech recognition," in *Spoken Language Technologies for Under-Resourced Languages*, 2012.

[24] M. Y. Tachbelie, S. T. Abate, and L. Besacier, "Using different acoustic, lexical and language modeling units for asr of an under-resourced language–amharic," *Speech Communication*, vol. 56, pp. 181–194, 2014.

[25] S. A. Shanmugam, "A hybrid approach to segmentation of speech using signal processing cues and Hidden Markov Models," M. S. Thesis, Department of Computer Science Engineering, IIT Madras, India, July 2015. [Online]. Available: "http://lantana.tenet.res.in/thesis.php"

[26] J. Lim, "Spectral root homomorphic deconvolution system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 223–233, 1979.

[27] T. Nagarajan, V. K. Prasad, and H. A. Murthy, "Minimum phase signal derived from root cepstrum," *Electronics Letters*, vol. 39, no. 12, pp. 941–942, 2003.

[28] Y. jun Kim and A. Conkie, "Automatic segmentation combining an hmm-based approach and spectral boundary correction," in *ICSLP*, 2002, pp. 145–148.

[29] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[30] A. Baby, N. L. Nishanthi, A. L. Thomas, and H. A. Murthy, "A unified parser for developing Indian language text to speech synthesizers," in *International Conference on Text, Speech and Dialogue*, Sep 2016, pp. 514–521.

[31] S. Young and S. Young, "The htk hidden markov model toolkit: Design and philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[32] A. Stolcke, "Srilm – an extensible language modeling toolkit," in *IN PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP 2002*, 2002, pp. 901–904.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.