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

Segmental and Supra-Segmental Feature Based Speech Recognition System for Under Resourced Languages

Tanmay Bhowmik[1], *Shyamal Kumar Das Mandal*[2]

[1]School of Computer Science,
University of Petroleum and Energy Studies (UPES),
Energy Acres Building, Bidholi, Dehradun - 248007, Uttarakhand, India
[2]Centre for Educational Technology, IIT Kharagpur
tanmay.bhowmik@ddn.upes.ac.in, sdasmandal@cet.iitkgp.ernet.in

Abstract

In detection-based, bottom-up speech recognition procedures, the segmental features like phonological feature based speech attributes act as one of the key component for the recognition model. In this study, place and manner of articulation based phonological features have been detected and they are integrated with the supra-segmental parameters of speech to develop the Auotmatic Speech Recognition (ASR) system for various under-resourced languages. For detection purpose a bank of phonological feature detector has been designed. Deep Neural Network (DNN) based attribute detector performed well to detect the phonological features. This paper also reports a comparative distribution of the (DNN) based attribute detector and the same using multi layer Perceptron (MLP). For continuous spoken speech, the Bengali CDAC speech corpus has been used. The deep neural based attribute detector achieved an average frame level accuracy of 88.26% is achieved whereas the same for MLP based detector is measured as 86.18%.

Index Terms: Segmental feature, Supra segmental feature, Bengali phoneme recognition, Deep Neural Network, Automatic Speech Recognition, Phoneme confusion

1. Introduction

In modern days, most of the people expect to have spoken information exchange with computers. So they need to have a communication media in speech mode between human and computers as speech is the most convenient and powerful way of communication.

Spoken language technology is very essential for the common people to have an easy communication with computers. Since speech is the most convenient and powerful way of communication, it is necessary to have a communication media in speech mode between man and machine. People expect to have a spoken information exchange with a computer.

Further, spoken language communication with computers provides a hand-free operation. This can create a good impact to have direct access to e-knowledge for more and more numbers of people. Acquisition of human speech represents the highest form of human cognition. The ability to have a clear conversation between a computer and a human is always an open challenge. This can be achieved by recognition and synthesis of speech.

State-of-the-art automatic speech recognition systems are based on the statistical approach [1] which uses a pattern matching framework. In that structure spoken utterances are represented as stochastic patterns. Data driven techniques are employed here. More the data will be employed, better performance will be delivered by this type of Automatic Speech Recognition (ASR) system. In general, top-down approaches are adopted for all the constrictions to be represented in a compact finite state network, collected by HMM states. Maximum a posteriori procedure is used to find the most possible match for a given speech segment and that most possible sequence of words in the FSN is identified as recognized sentence [2]. So to perform better, it needs to collect more and more data. This kind of top-down search technique achieved remarkable success in the field of ASR. However, the recognition accuracy falls drastically in case of spontaneous speech and performance level cannot match with the level of human speech recognition (HSR).

The statistical method of speech recognition is also depends on the lexicon which is mostly built based on written language. But to get more closer result as HSR, the conventional ASR technologies should be built based on spoken language lexicon instead of written language lexicon. But there are mere availability of spoken language lexicon for some under-resourced languages like Bengali, Oriya, Assamese, etc. So the need for developing the spoken language corpus for under-resourced languages are increasing day-by-day. And to develop the ASR system for these languages the need for applying the bottom-up techniques of speech recognition [3] is also gradually gaining much importance in speech recognition community. In this type of technique the phonological features that is the place and manner of articulation of the phonemes are detected first and they are integrated to identify the phonemes. The phonemes are considered as the smallest unit of speech. Detection of place and manner of articulation leads to the recognition of phonemes.

Speech recognition accuracy for isolated word recognition, read speech corpus like news broadcast, etc. were found as more than 95% using the conventional speech recognition techniques. But the recognition accuracy extremely decreases in case of spontaneous speech recognition. The robustness of a speech recognition system broadly depends on the performance on spontaneous speech. Actually in state-of-the-art recognition systems various knowledge sources of the speech events are not directly incorporated into the recognition model. That is why the ASR systems could not match with the performance level of human speech recognition (HSR). Speech scientists started to explore new techniques including sentence and word boundary detection, pronunciation modeling, adaptation of language and acoustic model, etc. [4, 5]. They also tried to utilize the prosodic knowledge sources [6] and the rich set of information

10.21437/SLTU.2018-47

which is embedded in speech knowledge hierarchy beyond the transcription of spoken utterances [7].

In this research work, first the places and manners of articulation are detected for Bengali language. For this purpose a deep neural network (DNN) based detection model has been generated. The results has been compared with the results obtained by the baseline system. The baseline system has been developed by multi layer perceptron based model. Then these features are integrated with the supra-segmental parameter of speech, speech prosody to develop a knowledge-based ASR system.

2. Segmental and Supra-segmental parameters of Speech

In this research work the study has been carried out to identify place and manner of articulation. So this investigation has been performed at the phonetic and phonological level. That is why this work is performed in segmental level as each phoneme is considered to be one segment of speech. Once the investigation is moved on to look larger chunks of speech, which span a group of segments such as a complete word or phrase, the dealing of features will be carried out in supra-segmental level. In this research, the prosodic feature has been considered as supra-segmental feature.

3. Knowledge-based Modeling of ASR System

Speech scientists are trying to integrate different knowledge sources into the state-of-the-art ASR systems during last few years to improve the ASR performance. The sound-specific feature like voice onset time (VOT) measurement was proved as more powerful than the spectral features to discriminate the stop consonants t̪ and d̪ [8]. This procedure also reduce a 50% error rate in comparisosn to state-of-the-art model. Keyword spotting was another crucial idea to incorporate in conventional ASR system as human being is a good listener to identify the keyword from any kind of spontaneous speech. Combined approach of key-word spotting and utterance verification produced better result in case of recognition from unstructured continuous-spoken speech data [9]. The artificial neural network was used to produce some speech attributes which were used to design a large vocabulary continuous speech recognition (LVCSR) module as a knowledge-based front-end system [10]. The knowledge-based features were used for training of a group of HMMs. A ROVER system was designed using the combination of 44 phoneme features, 60 speech attributes, and the Mel Frequency Cepstral Coefficient (MFCC) features [11]. This system obtained 20% over the result produced by the best baseline system. A popular feature-based speech recognizer was built using the knowledge-based representation with Dynamic Bayesian Network [12]. Recognition of articulatory features was proposed by Frankel and King [13]. The articulatory features were detected by the HMM/ANN combined model also [14] and combination of acoustic and articulatory features were used for robust speech recognition [15].

The ASR performance does not match with the performance level of HSR in case of continuous speech recognition. As a result speech scientists need to incorporate some of the speech related features in the recognition model. The ASR model needs to be enlarged by embedding the related information which remain unobserved in the speech utterances [7].

4. Speech Material

The speech material has been used for this study is in Bengali language. Bengali is one of the more important Indo-Aryan (IA) languages. It is the official state language of the Eastern Indian state West Bengal and the national language of Bangladesh. Bengali is one of the most spoken languages (ranking fifth) in the world with nearly 250 million total speakers [16]. In India, most of the Bengali-speaking population are found in West Bengali (85%), Tripura (67%), Jharkhand (40%), Assam (34%), Andaman & Nicobar Islands (26%), Arunachal Pradesh (10%), Mizoram (9%), and Meghalay (8%) [17]. Dialect wise Bengali language is divided into two main branches; eastern and western. Rarha (South), Varendra (North Central) and Kamrupa (North Bengal) dialect clusters are in western branch. Rarha is further sub-divided into South Western Bengali (SWB) and the Standard Colloquial Bengali (SCB) which is spoken around Kolkata [18]. The present study is based on Standard Colloquial Bengali (SCB).

5. Experimental Setup

5.1. Speech corpus

The speech corpus from Centre for Development of Advanced Computing (CDAC), India [19], has been used for continuous spoken Bengali speech data. This is a high quality speech corpus labeled at both the phone level and the word level. A random subset of the corpus consist of 12 types of sentences including Complex affirmative, Complex negative, Simple affirmative with a verb, Simple affirmative without a verb, Simple negative, Compound affirmative, Compound Negative, Exclamatory, Imperative, Passive, WH questions, and Yes-No questions have been used for this experiment. 500 sentences were taken for training and 200 sentences for testing. A small development set consists of 50 sentences was split off from the training set to decide when to stop the MLP training. CDAC speech corpus contains four female, and five male speakers voice of different age group varies from 25 to 40. Audio and text were time-aligned. The sampling frequency of all the recording sentences in the speech corpus is 22050 Hz. All the recordings were re-sampled to 16000 Hz. Resampling reduces the irrelevant information that is not produced by human voice, in the sound files [20][21]. All the experiments were gone through a subset of TIMIT corpus also. The TIMIT subset contains randomly chosen 500 sentences of all the types that exist in the corpus [22]. Those sentences are spoken by 50 speakers, including 25 male speakers, and 25 female speakers of different age group vary from 20 to 50 years.

5.2. Baseline systems

A phoneme recognition system using MLP was built for baseline approach. It was a Three layered MLP based system with 500 hidden units in the hidden layer and the posterior probability for every phoneme was derived in the output layer. The whole training set with 500 sentences were used to train the system and tested on 200 sentences.

5.3. Input features

For input data of this system we adopt Mel-Frequency Cepstrum Coefficients (MFCC) features among all possible parametric representations of the speech signal. 12th order MFCC features along with log energy and the 0th cepstral coefficient is computed for each frame. The derivative and double-derivative

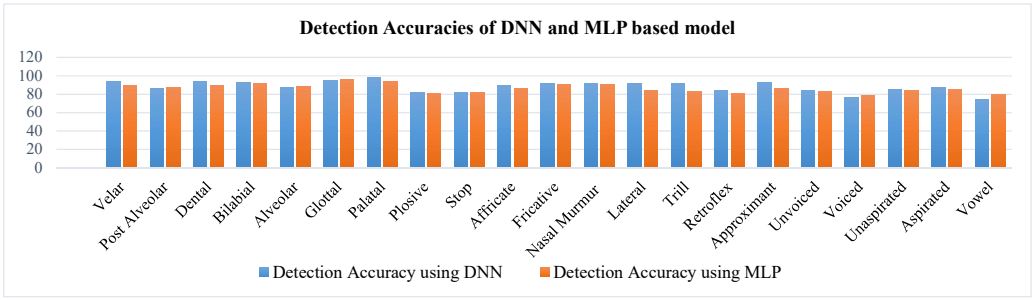

Figure 1: Distribution of Detection Accuracies for DNN and MLP based model

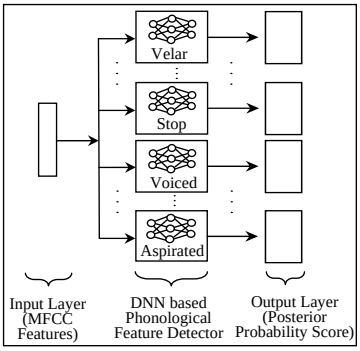

Figure 2: DNN-based phonological feature detection model

is also computed for the 13 MFCC features using the equations $\delta 1[y(n)] = [y(n+1) - y(n-1)]$ and $\delta 2[y(n)] = [0.5y(n+1) - 2y(n) + 0.5y(n-1)]$ respectively to yield a 39 dimensions input feature vector. A contextual representation of 11 frames of 39 dimensions input feature vector has been taken to prevent data loss; five frames looking back and five frames looking ahead. In the absence of a speech frame, zero is appended to complete the context. This creates the input node of size $39 \times 11 = 429$.

5.4. Softwares

For the MLP based system the Matlab Neural Network toolkit has been used. A deep learning toolkit [23] has been used for phoneme recognition using the Stacked Denoising Autoencoder. For the analysis of prosodic features it was required to extract the fundamental frequencies. PRAAT was used for this purpose [24].

5.5. Methodology

5.5.1. Architecture of MLP based recognizer

A Three layered MLP has been used here with 500 units in the hidden layer for the CDAC speech corpus. In input layer the acoustic features i.e. the 39 dimensions input feature vector consists of MFCC coefficients are given. A contextual representation of 11 frames having 39 dimensions input feature vector per frame has been taken to prevent data loss; five frames looking back and five frames looking ahead. In the absence of a speech frame, zero is appended to complete the context. This creates the input node of size 39 × 11 = 429. In output layer, the posterior probability distributions are derived for each phoneme.

5.5.2. Architecture of DNN based recognizer

For the deep architecture, during pre-training, denoising autoencoders are trained. In this experiment, two autoencoders are stacked to form the deep architecture. For each autoencoder the sigmoid function has been used as the non-linear activation function. 300 hidden units are used in each hidden layer. There are 48 phonetic classes are used in the CDAC speech corpus. So in output layer there will be 48 nodes. The learning rate is kept as one and the mini batch size is fixed at 120. The input zero masked fraction value is taken as 0.5. In input and output layer, there are the MFCC features and the phone posterior probabilities respectively. Basic block diagram of the DNN-based model is given in Figure 2.

5.5.3. Analysis of Supra-segmental Parameters

In the analysis phase of supra-segmental parameters, the continuous speech signal is broken into prosodic word candidates based on the supra-segmental parameters (F_0 contour). The automatic extraction of accent command which is used in prosody modeling has been carried out to detect the prosodic word boundaries in continuous speech. The prosodic word boundary is detected for Bengali continuous speech based on the fundamental frequency (F_0) contour analysis of Bengali continuous speech signal. Bengali is a bound-stressed language where stress is always found on the first syllable [25]. Due to the stress on the first syllable, the word-level F_0 value varies from low to high [26]. F_0 contour exhibited a significant role in manifesting the prosodic information of an utterance [27]. A prosodic word of Bengali can be defined by the presence of negative accent command at the beginning of a word [26]. The logarithmic, continuous F_0 contour is decomposed into multiple Intrinsic Mode Functions (IMF) by the Empirical Mode Decomposition (EMD) method. The EMD method decomposes the F_0 contour into very rapidly varying components or IMF that correspond to micro-prosody, rapidly varying components with the local variation which represents the accent components, slowly-varying components with the global variation that represents phrase components, and the residual components. The procedure of EMD analysis along with the accent and phrase component extraction procedure is shown in Figure 3.

6. Results and Discussions

In this experiment, the average phoneme recognition accuracy is obtained as 88.26% from the deep structure based model, while the MLP-based baseline system deliver the average recognition accuracy as 86.18%. The comparative distribution of the accuracies obtained by these two models in Figure 1.

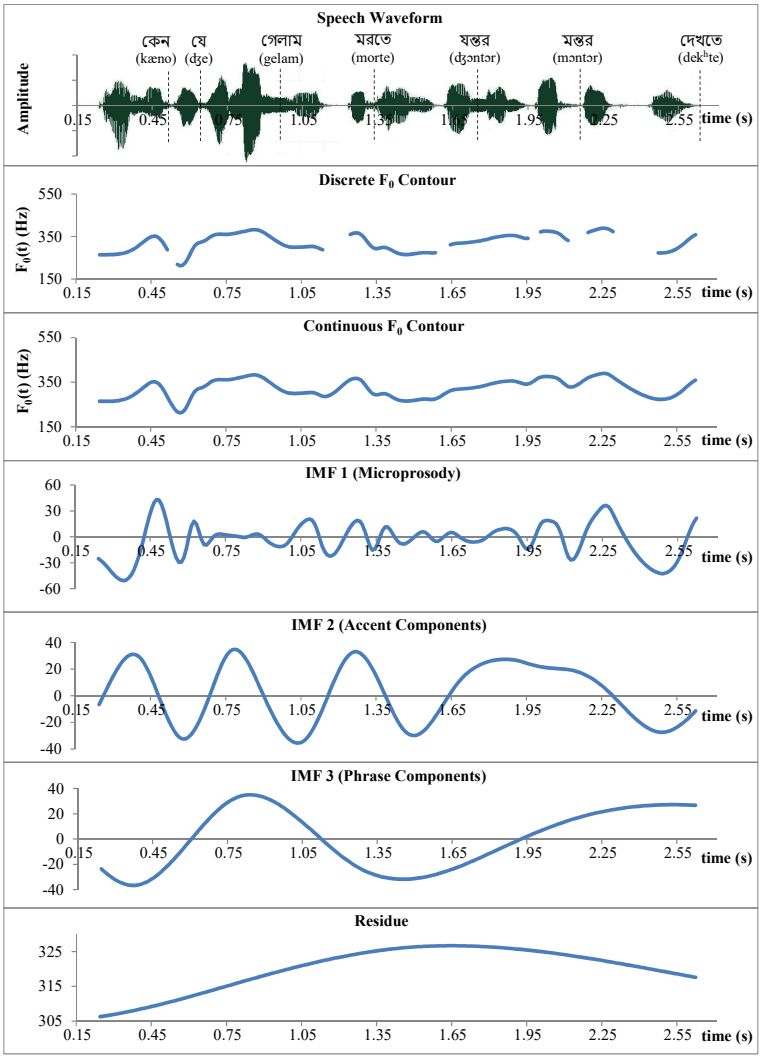

Figure 3: Empirical Mode Decomposition of Fundamental Frequency (F_0) Contour for Bengali Continuous Speech

Detection of word boundaries in continuous speech is a tedious process, due to the absence of a definite pause or silence in the word boundary position. However, the prosodic parameters of continuous speech indicate prosodic word boundaries which may not be identical with the lexical word (written word) boundaries. This work proposes a method for detecting such prosodic word boundaries in Bengali continuous speech based on different prosodic parameters. Bengali is a bound-stressed language, and the stress is always on the first syllable of a prosodic word. Empirical Mode Decomposition (EMD) of the logarithmic fundamental frequency (F_0) contour of Bengali continuous speech is used for detecting the prosodic word boundaries. In Figure 3 an example of EMD analysis is given. The logarithmic F_0 contour is decomposed into microprosy, accent, phrase and residual components. 200 Bengali readout sentences, read by ten speakers, are analyzed for this study. An overall prosodic boundary detection accuracy of 92.07% is achieved with 90.73% precision and 88.31% recall. So using this procedure a lexicon with prosodic words can be developed.

Using the classification technique of phoneme based on manner of articulation [28], the phonemes can be classified into several groups. The robustly identified groups of phonemes are used for manner based labeling of the prosodic words. This creates a pseudo-word dictionary. A single pseudo-word may consist of one or more prosodic words. Those groups of prosodic words are defined as Cohort. Using the Cohort Analysis the final recognition accuracy can be achieved. Unique cohorts are directly recognized as the corresponding prosodic words. A lexical expert system [29] is used to separate the cohorts with size two or more.

7. Conclusion

The spoken word dictionary is not widely available for various under-resourced languages. So the development of prosodic word dictionary is very much fruitful for speech recognition area for under-resourced languages. The performance of the proposed combined segmental supra-segmental model can be improved with a larger prosodic word lexicon. The task of generation of prosodic word lexicon is an ongoing task. This model can be used also for other under-resourced languages which are bound-stressed.

8. References

[1] F. Jelinek, "Speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, pp. 532–556, 1976.

[2] L. 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.

[3] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "A bottom-up stepwise knowledge-integration approach to large vocabulary continuous speech recognition using weighted finite state machines," in *INTERSPEECH, Florence, Italy*, 2011, pp. 901–904.

[4] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[5] S. Furui, "Recent progress in corpus-based spontaneous speech recognition," *IEICE transactions on information and systems*, vol. 88, no. 3, pp. 366–375, 2005.

[6] D. Vergyri, A. Stolcke, V. R. R. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–208.

[7] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proc. ICSLP*, vol. 4, 2004.

[8] P. Niyogi, P. Mitra, and M. M. Sondhi, "A detection framework for locating phonetic events." in *ICSLP*, 1998.

[9] C.-h. Lee and R. A. Sukkar, "Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition," Oct. 7 1997, uS Patent 5,675,706.

[10] B. Launay, O. Siohan, A. Surendran, and C.-H. Lee, "Towards knowledge-based features for hmm based large vocabulary automatic speech recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I813–I817.

[11] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[12] K. Livescu, J. R. Glass, and J. A. Bilmes, "Hidden feature models for speech recognition using dynamic bayesian networks." in *INTERSPEECH*. Citeseer, 2003.

[13] J. Frankel and S. King, "A hybrid ann/dbn approach to articulatory feature recognition," 2005.

[14] K. Kirchhoff *et al.*, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *ICSLP*. Citeseer, 1998, pp. 891–894.

[15] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3, pp. 303–319, 2002.

[16] M. P. Lewis, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the world*. SIL international Dallas, TX, 2016, vol. 19.

[17] Wikipedia, "States of india by bengali speakers — wikipedia, the free encyclopedia," 2015, [Online; accessed 23-March-2016]. [Online]. Available: \url{https://en.wikipedia.org/w/index.php?title=States_of_India_by_Bengali_speakers&oldid=676468584}

[18] K. Bhattacharya, *Bengali Phonetic Reader*. Central Institute of Indian Languages, 1988, vol. 28.

[19] S. D. Mandal, A. Saha, and A. Datta, "Annotated speech corpora development in indian languages," *Vishwa Bharat*, vol. 6, pp. 49–64, 2005.

[20] D. Feinberg, B. Jones, A. Little, D. Burt, and D. Perrett, "Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices," *Animal Behaviour*, vol. 69, no. 3, pp. 561–568, 2005.

[21] P. Ladefoged, *Elements of acoustic phonetics*. University of Chicago Press, 1996.

[22] J. Garofolo, L. D. Consortium *et al.*, *TIMIT: acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium, Philadelphia, 1993.

[23] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data," Master's thesis, 2012.

[24] P. Boersma and D. Weenink, "Praat software," *Amsterdam: University of Amsterdam*, 2013.

[25] B. Hayes and A. Lahiri, "Bengali intonational phonology," *Natural Language & Linguistic Theory*, vol. 9, no. 1, pp. 47–96, 1991.

[26] S. D. Mandal, A. H. Warsi, T. Basu, K. Hirose, and H. Fujisaki, "Analysis and synthesis of f0 contours for bangla readout speech," in *Proc. of Oriental COCOSDA*, 2010.

[27] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in *Speech Prosody 2004, International Conference*, 2004.

[28] T. Bhowmik and S. K. D. Mandal, "Manner of articulation based bengali phoneme classification," *International Journal of Speech Technology*, vol. 21, no. 2, pp. 233–250, 2018.

[29] S. Das Mandal, "Role of shape parameters in speech recognition: A study on standard colloquial bengali (scb)," Ph.D. dissertation, 2007.