Development of Assamese Text-to-speech System using Deep Neural Network

Abhash Deka*, Priyankoo Sarmah*, Samudravijaya K†, S R M Prasanna†

*Center for Linguistic Science and Technology
Indian Institute of Technology Guwahati, Guwahati, Assam 781039 INDIA
Email: {abhash, priyankoo, samudravijaya}@iitg.ac.in
†Department of Electronics and Electrical Engineering
Indian Institute of Technology Dharwad, Dharwad, Karnataka 580011 INDIA
Email: prasanna@iitdh.ac.in

Abstract—This paper describes the development of a text-to-speech system for Assamese language, using Deep Neural Network (DNN). The system is trained with speech data, collected by a consortium, that is available free of cost for academic use. The DNN based method eliminates the need for a grapheme to phoneme conversion; rather, it synthesizes speech directly from the UTF-8 based Assamese script. The results of objective and subjective evaluations confirm that the Assamese speech synthesized using DNN approach is better than the ones synthesized using the traditional hidden Markov model based text-to-speech system.

I. INTRODUCTION

Development of text-to-speech (TTS) systems in Indian languages received a fillip from the Indian languages TTS consortium project sponsored by MeitY, Government of India [1]. The efforts of the TTS consortium resulted in the development of TTS systems in 13 Indian languages, namely, Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, Gujarati, Odia, Assamese, Manipuri, Kannada, Bodo and Rajasthani [2].

Initial versions of the TTS systems followed concatenative or unit selection synthesis approaches, which use small units of speech sounds to produce synthetic speech [2]. However, to produce all speech sounds in a variety of acoustic-phonetic and prosodic contexts, a large speech database containing all possible phonemes and their combinations is required to be stored within the machine. This renders the task difficult for a low-resource language such as Assamese. However, by using statistical parametric speech synthesis methods, such as the Hidden Markov Model (HMM), the need to store large amount of speech sound files is obviated. Rather, this approach stores parameters, such as fundamental frequency and cepstral coefficients, of statistical models of speech sounds in order to synthesize speech corresponding to the input text. This method is superior to the concatenation based approach [3] as it has low requirement of training data, useful for low resource languages. Moreover, this approach has the ability to change voice characteristics, such as the speaking style of the speaker and speaker emotions. However, in this method, the quality of synthesized speech is not as natural as that of the concatenative approaches, requiring additional processing of the synthesized speech. Zen et al. described three reasons that are responsible for degraded speech: vocoding, accuracy of acoustic models and over-smoothing [4].

To reduce such degradation in speech quality, and to improve the naturalness of synthesized speech, Deep Neural Network (DNN) based approaches have been used in recent years. Availability of large datasets and improvement in computational power has made the use of DNN for TTS easier. Also, recent studies show that building TTS systems using deep neural network results in more natural synthesized speech [5]. This is due to the ability of the DNN to model long-span frames [6], and to model high dimensional and strongly correlated features, and to find a highly non linear mapping between input and output features. An investigation revealed that primary reasons for improved naturalness ratings of humans listening to synthetic speech produced by TTS systems using DNNs were the replacement of decision trees with DNNs and moving from state-level to frame-level predictions [7].

Most of the 13 language TTS systems, implemented by TTS consortium, were initially developed using the festvox framework [2], [8]. Hidden Markov model based text-to-speech (HTS) engines were developed subsequently [3]. The development of HTS systems was facilitated by TBT, an open source toolkit to build multiple language TTS systems [9]. A by-product of the TTS consortium project was the creation of a speech database in many Indian languages for implementing TTS systems [1]. We used the Assamese language module of this speech database to build a new TTS system using DNN to model the mapping between text based features and the corresponding acoustic features needed for generation of speech signal. The current system generates speech waveform corresponding to an input Assamese text in UTF-8 format. The quality of the speech generated by this DNN based system is better than those generated by TTS systems following the conventional statistical parametric approach using hidden Markov model. Here, we report the development of a DNN based Assamese TTS system, and subjective as well as objective evaluation of the quality of speech synthesized by the system.

The rest of the paper is organized as follows. The speech database and the toolkit used for implementing the TTS system are described in Section II. The details of the subjective and objective evaluation methodologies are also given in the same
In this section, we describe the speech data and the transcriptions used to train the DNN based Assamese TTS system. The details of the implementation of the TTS system, and methodology of conducting objective as well as subjective evaluation tests are also given in this section.

A. Database

To build our text-to-speech synthesis system for Assamese language, we used the speech database of the TTS consortium. The speech data can be downloaded for academic and research use [10]. The Assamese database contains sound files generated by a male and a female speaker. We used 8941 audio files, spoken by a male speaker, to build the TTS system. Each sound file contains one spoken sentence. The digitized speech is stored in raw format (48000 Hz, signed 16-bit PCM, little endian, mono). The total duration of recorded speech data is 12.95 hours. A single text file contains the text corpus comprising of 8,941 sentences. We created 8,941 text files, each containing the sentence corresponding to one sound file. The text corpus contains 32,136 unique words [10]. The speech files and the corresponding transcriptions were used to build the Assamese text-to-speech system using DNNs as described below.

B. Implementation of TTS system based on DNN

Figure 1 shows a block diagram of the Assamese TTS system. It contains 4 processing blocks: (i) a text processor to generate linguistic specification of the input Assamese text, (ii) a (duration) DNN to generate duration information of the textual unit from its linguistic specification, (iii) another (acoustic) DNN to map the linguistic features to acoustic features, and (iv) a vocoder to generate speech data from the acoustic features and duration information of the textual unit. A brief description of the task carried out by each processing block as well as the software tools used to carry out the tasks is provided below. For details of application of DNNs for TTS, one may read a tutorial presented at Interspeech 2017 [11].

1) Text Analysis: In order to derive linguistic features from Assamese text in UTF-8 format, we have used Ossian [12] as a front end. Ossian is a collection of python codes that aids in building TTS systems. It is an open source toolkit, distributed using Apache License 2.0. Ossian supports the use of neural nets trained with the Merlin toolkit [13] as duration and acoustic models. Ossian relies on the HMM toolkit (HTK) [14] and HMM based text-to-speech (HTS) system [15] for alignment of speech data with transcription. The biggest advantage of Ossian is that it does not require any language specific knowledge to extract the linguistic features. The UTF-8 characters are used to tokenise the text, and characterise tokens as words, white space, punctuation etc. Ossian uses a letter/grapheme based approach in which the names of letters are used directly as the names of speech modelling units in contrast to the traditional approach where phonemes are used as the speech units. This approach eliminates the need to have any language specific knowledge such as phonetic categories (vowel, nasal, approximant etc.) as well as part of speech categories (noun, verb, adjective etc.).

Using Ossian, the following speech features were extracted from the Assamese speech data: mel-generalized cepstrum, logarithm of fundamental frequency and mean band aperiodicity. The output is formatted as HTS-style labels using state-level alignment. The labels are then converted into sequence of vectors of binary and continuous features by employing HTS-style questions to derive the features from the label sequence. The resulting linguistic features are used as input features for training duration and acoustic DNNs.

2) Deep Neural Networks: The current TTS system takes linguistic features as input, and employs a combination of two DNNs to predict acoustic features, which are then passed to a vocoder to produce the speech waveform. One (duration) DNN was trained to predict the duration of letter units from acoustic features. Another (acoustic) DNN with 6 hidden layers was trained to map the input linguistic features and associated duration features into acoustic features.

The sequence of vectors of binary and numeric features generated by the text processor were normalized in the range of [0.001, 0.99] before feeding to the input layer of DNN. The duration DNN was trained using the (forced Viterbi) time-aligned data, to predict duration of a state of a HMM or the duration of an entire HMM representing a quin-letter (a letter with 2 left and 2 right letter contexts).

The acoustic features at the output layer of acoustic DNN are Line Spectral Pairs (LSP), Fundamental frequency (F0) and unvoiced/voiced (U/V). The DNN was trained using Stochastic Gradient Descent algorithm, with an initial value of the learning rate as 0.02. The weights of DNN were estimated by using frame-aligned pairs of input and output features, extracted from the training data, to minimize errors between
outputs mapped by the DNN and the target outputs [6]. The sequence of output features were normalized to have zero mean and unit variance.

During synthesis, duration of a quin-letter is predicted first. Then, the linguistic features of the quin-letter and its predicted duration are fed to the acoustic DNN to predict the sequence of acoustic feature vectors. Thanks to the duration prediction, typically many acoustic feature vectors get generated by the acoustic DNN corresponding to one input linguistic feature vector.

3) Vocoder: To synthesize the time waveform using the acoustic features estimated by the DNN, we have used WORLD [16], a free vocoder. Merlin toolkit contains a version of WORLD, modified to satisfy the requirements of Merlin.

By setting the predicted output features from the DNN as mean vectors, and using the pre-computed (global) variances of output features from all training data, the speech feature generation module generates smooth trajectories of speech parameter features which satisfy the statistics of static and dynamic features. In WORLD, the vocal cord vibration is calculated on the basis of the convolution of the excitation signal with the minimum phase response of the spectral envelope, interpolated at excitation pulse locations along the time axis [16]. The F0 information is used to determine the temporal positions of the pulse location (origin of each vocal cord vibration).

C. Objective and Subjective Evaluation

In order to assess the expected improvement in quality of speech synthesized by the current system that uses DNN instead of GMM-HMM [8], we carried out both objective and subjective evaluations. For objective evaluation, we adopted the Perceptual Evaluation of Speech Quality (PESQ) [17], [18] as a measure of quality of synthetic speech. For objective evaluation, 100 sentences were randomly chosen from the database of 8941 Assamese sentences. The original Assamese sentences, spoken by a human subject, were the reference items; the corresponding synthesized sentences, using GMM-HMM [8] and DNN approaches, constituted the test set.

Subjective evaluation of the synthesized speech was conducted by 26 human raters, who are native speakers of Assamese. The Assamese speakers performed a Differential Mean Opinion Score (DMOS) task where they were asked to score the quality of the speech generated by a TTS system with reference to the speech of the same text spoken by an Assamese speaker [19]. In this Degradation Category Rating method, a subject listens to the utterance of a sentence as produced by a human followed by the speech of the same sentence synthesized by a TTS system, and rates the relative score of the synthetic speech. In the current work, audio files listened to by human raters, contained natural speech, a 440Hz beep and the corresponding synthetic speech in that order. The sampling frequency of the mono channel audio file was 48kHz.

Of the 100 sentences chosen for the objective analysis, 15 sentences were selected for the subjective evaluation. Speech data corresponding to each of these sentences was synthesized by both GMM-HMM and DNN based TTS systems. This yields 30 synthetic stimuli. In the evaluation task, a synthetic stimulus was played after playing the corresponding reference/natural speech. Each stimulus was presented twice to each and every subject, resulting in 60 total stimuli, per subject. The 60 stimuli were presented in random order for evaluation.

Each subject was instructed to rate the degradation of the synthesized speech in comparison to the human speech on a scale from 5 to 1, corresponding to the following judgments respectively: imperceptible, perceptible but not annoying, slightly annoying, annoying and very annoying. The methodology followed was in accordance with the guidelines provided in ITU–T Rec. P.913 [20] and discussed in detail by Pinson and Janowski [21].

The perceptual evaluation test was conducted via a Praat [22] based graphical user interface on a laptop computer. The reference and the test speech were normalized to have equal intensity. The subjects listened to the stimuli using a headphone with a flat response in the frequency range 20–20000Hz. The subjects were allowed to listen to the sentences as many times as they liked, before making a judgment. The graphical user interface allowed the subjects to click one of the five buttons to choose the degradation category they deemed appropriate. Each subject took about 20 minutes to complete the DMOS test. The evaluation scores were extracted and tabulated for statistical analysis as detailed in Section III-B.

III. RESULTS AND DISCUSSION

The results of objective and subjective evaluation of the quality of speech generated by the traditional GMM-HMM based TTS system and the current, DNN-based TTS system are presented and discussed in this section.

A. Objective evaluation

The standard objective measure of speech quality, PESQ [18], provides two scores: PESQ-MOS and PESQ-LQO (Mean Opinion Score; Listening Quality Objective). A mapping function maps the PESQ-MOS raw values in the range [-0.5, 4.5] to PESQ-LQO in the range [1, 5]. These two scores for speech synthesized by the two (GMM-HMM and DNN) TTS systems, are tabulated in Table I.

<table>
<thead>
<tr>
<th>TABLE I</th>
</tr>
</thead>
<tbody>
<tr>
<td>PESQ SCORES OF SPEECH SYNTHESIZED BY TTS SYSTEMS BASED ON GMM-HMM AND DNN</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>GMM-HMM</td>
</tr>
<tr>
<td>DNN</td>
</tr>
</tbody>
</table>

Since the difference between the average PESQ scores are marginal, we decided to conduct two separate paired t-tests for PESQ-MOS and PESQ-LQO values. There was a statistically significant difference in the PESQ-MOS scores of the synthesized speech using GMM-HMM and DNN approaches, with t(99)=8.4, p< 0.0001. Similarly, the two approaches
differed significantly with respect to PESQ-LQO scores; the corresponding $t(99) = 5.7$, $p < 0.0001$.

The PESQ-MOS scores of 100 synthetic speech signals are shown in the form of a bar chart in Fig. 2. One can see that the scores of the speech generated by the DNN system (blue bars) are, in general, higher than those of the GMM-HMM system (red bars). Similar trend is visible in the case of PESQ-LQO scores as shown in Fig. 3. In the case of both scores, 83 out of 100 synthesized sentences using the DNN approach have received higher PESQ-MOS and PESQ-LQO scores than the ones synthesized using the GMM-HMM approach.

The PESQ-MOS scores of 100 synthetic speech signals are shown in the form of a bar chart in Fig. 2. One can see that the scores of the speech generated by the DNN system (blue bars) are, in general, higher than those of the GMM-HMM system (red bars). Similar trend is visible in the case of PESQ-LQO scores as shown in Fig. 3. In the case of both scores, 83 out of 100 synthesized sentences using the DNN approach have received higher PESQ-MOS and PESQ-LQO scores than the ones synthesized using the GMM-HMM approach.

### Table II: Average of DMOS Scores of the Two TTS Systems

<table>
<thead>
<tr>
<th></th>
<th>DMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM-HMM</td>
<td>2.7</td>
</tr>
<tr>
<td>DNN</td>
<td>3.7</td>
</tr>
</tbody>
</table>

In this paper, we reported the development of an Assamese text to speech system that uses a deep neural network to map linguistic features of input text to acoustic features that was used by a vocoder to generate good quality speech signal. The Assamese TTS system was trained using the Assamese language module of the speech database of the TTS consortium. Both subjective and objective tests reveal that the quality of the speech synthesized by the DNN based TTS system developed by us is distinctively better than the quality of speech synthesized by the GMM-HMM based TTS system. Our system also obviates the need for grapheme to phoneme conversion, and generates speech corresponding to the text written in Assamese script in UTF-8 format.

B. Subjective evaluation

A subjective evaluation of the synthesized stimuli, using DMOS scores as described in Section II-C, showed that all 26 speakers consistently rated the synthesized stimuli generated by the DNN approach as better. The average of DMOS scores corresponding to the GMM-HMM and DNN TTS systems are are provided in Table II. The average DMOS score of DNN based TTS system is 1.0 higher than that of the conventional GMM-HMM based TTS system.

Each of the 15 synthetic speech was scored by 26 native Assamese speakers. The average of DMOS scores of all 26 listeners for each of the 15 stimuli generated by GMM-HMM and DNN methods are presented in a stepped chart in Fig. 4. As seen in the figure, in the case of all 15 stimuli, human listeners have rated the quality of the DNN generated stimuli better than the GMM-HMM generated stimuli. A paired t-test confirmed that the DMOS scores obtained by the DNN generated stimuli are significantly different from the ones generated by the GMM-HMM approach [ $t(779) = 24.7$, $p < 0.0001$].

IV. Conclusion

In this paper, we reported the development of an Assamese text to speech system that uses a deep neural network to map linguistic features of input text to acoustic features that was used by a vocoder to generate good quality speech signal. The Assamese TTS system was trained using the Assamese language module of the speech database of the TTS consortium. Both subjective and objective tests reveal that the quality of the speech synthesized by the DNN based TTS system developed by us is distinctively better than the quality of speech synthesized by the GMM-HMM based TTS system. Our system also obviates the need for grapheme to phoneme conversion, and generates speech corresponding to the text written in Assamese script in UTF-8 format.

ACKNOWLEDGMENT

This work is part of a project titled “Development of Speech Interface for Form-filling application (SiFA) in Five Indian languages”, funded by the Ministry of Electronics and Information Technology (MeitY) and the Ministry of Human Resource Development (HRD), Government of India, under the IMPRINT research grant program. The implementation of the Assamese TTS system was benefited by the Assamese speech data recorded as part of the TTS consortium, sponsored...
by MeitY. The authors would like to express their gratitude to Mr. Giridhar Pamishetty of IIT Hyderabad for his help. The authors would also like to express their sincere gratitude towards the reviewers of NCC 2019 for their constructive suggestions.

REFERENCES


