Syllable based approach for text to speech synthesis of Assamese language: A review

Parismita Sarma and S K Sarma
Department of IT, Gauhati University, Guwahati-14, India
E-mail: parismita.sarma@gmail.com, sks001@gmail.com

Abstract. In this review article authors are trying to put light on text to speech synthesis of Assamese language using unit selection concatenative speech synthesis technique. Assamese is one of the North East Indian languages spoken by millions of people. This article tries to highlight some major difficulties when developing the synthesizer. The speech units used to concatenate for building the synthesizer is syllable. Assamese is a syllable centric language and syllable based concatenation gives more natural sound. A discussion is done with phoneme and diphone units as well. There is also a short overview on development process of the Assamese synthesizer on festival framework as a part of review. Another challenging task the researchers dealt with is building speech corpus for a low resource language like Assamese.

1. Introduction

Text to speech conversion is a system where input is text and output is the corresponding speech waveform. In most of the languages in the world including many Indian languages, already TTS (Text to Speech Synthesizer) are built. But most of them are lagging in natural sound, missing expression of the context when the machine pretend to speak.

There are different methodologies that are used to build speech synthesizers. Mostly used approaches are Articulatory, Formant and Concatenative speech synthesis[1]. In this paper researchers will concentrate on concatenative speech synthesis which is a basic technique and generally gives more natural and spontaneous output for most of the Indian languages. Assamese synthesizer using articulatory and formant synthesis are also developed by some researchers. But due to more natural output and less development complexity concatenative synthesis are mostly preferred for Indian languages. In concatenative speech synthesis units like phones, diphones or syllable are used as basic units of utterance that will be used to concatenate and stored as a dataset. Most appropriate units are searched using some algorithms and concatenate them after extracting from the dataset. A natural sounding speech is comprised of phrase break and silence that are used to give word gap. Stress, prosody, tune etc are important supra segmental characteristics of utterance that are considered at the time of synthesis[2]. Initial afford was made to store commonly used words of a language as a large database. This corpus based approach works for some sentences but major drawback noticed is failure to pronounce the proper nouns. This procedure of word concatenation failed to incorporate proper intonation to the utterance and phase break among the units. In the next phase diphones were considered for synthesis and this approach still works for many languages. Though word level concatenation were able to produce natural speech, synthesis flexibility was missing in this approach[3]. To balance these two issues concept of half phone was adopted which is named as
diphone. Phonemes are divided into two halves and joining is performed on these small parts. Diphone concatenation is characterized as follows\cite{4}.

1. Each diphone is recorded by a single speaker.
2. Diphones are cut from speech and stored in database.

Synthesis is done in the following way:
1. Corresponding diphones for the target phones are selected from the database.
2. At the boundary of the diphones some signal processing is done and they are concatenated.
3. Again signal processing is performed to transform prosody like f0, duration etc. on the diphone sequence to attain the required prosody.

Due to its capability to prevent coarticulation phenomenon, diphones are more preferred than phones for concatenative synthesis. Coarticulation is defined as a condition where isolated speech units are influenced by its predecessor and successor units and behave a bit differently\cite{4}. This phenomenon results in discontinuities at boundary level whereas in case of diphone coarticulation is merged into the units themselves.

1.1. Assamese language
In the state of Assam in India, millions of people speak Assamese and some other people who migrated to different places of the world. Indo-European family is root of Assamese language, other root is found at Indo-Iranian family. The three groups of Assamese language are Dardic, Indic as well as Iranian\cite{5}. Another name of Indo-Aryan family is Indic. There are eight numbers of vowel phonemes here among which three are nasalized vowel phonemes. Moreover fifteen diphthongs and twenty one consonants are found in Assamese language. Important characteristics of the language is that there is no Dental, Retroflex, Lateral fricative, Labiodental, Uvular and Pharyngeal phonemes\cite{6}. On the other hand alveolar, approximants, nasal-oral stops, fricative and laterals are mostly found in Assamese. Another important feature is recurrent use of velar nasal /ŋ/. This is a unique phoneme, other Indo-Aryan languages use some other homographic pronunciations for it. /x/ is the velar fricative present in Assamese language found nowhere in Indian language.

1.2. Concatenative synthesis
Now we are going to discuss about concatenative speech synthesis. This method of synthesis is done by combining already recorded speech units using specific mathematical and probabilistic methods. The synthesizer built by this methodology generally gives pleasant and natural speech. A huge speech database is needed for this and it consumes a vast memory space. Main challenge here is to determine the length of the units to be concatenated. For most of the Indian language syllable is taken as the basic unit to concatenate, as most of them are syllabic in nature\cite{7}. A brief discussion is written below on different units that are considered for concatenation.

1.2.1. Word concatenation
It is observed that in case of restricted domain like reservation system or announcement system word concatenated synthesizer gives better result. A number of words that are necessary for the system are recorded, extracted and concatenated at run time. The words recorded may be in normal as well as news reading mood. One advantage of this methodology is it may reduce memory spaces necessary for database but at the same time drop of coarticulation at the joining points is an issue. Other matters are pause duration, energy level of the consecutive words which must be compatible to each other for natural sound\cite{8}. Pre recorded words should be pronounced normally and only one speaker’s voice is preferred for word concatenated speech synthesis.

1.2.2. Phoneme concatenation
A total set of phonemes with all auditory features are necessary for phoneme concatenation. There are only fifty or sixty numbers of total phonemes in the world for all language\cite{9}. Generally some selective phonetically rich sentences are selected and recorded for this purpose. Then phonemes are extracted by any automated techniques, wave editor or by hand work. In the next phase individual units are normalized such that the synthesizer when speaks can provide appropriate meaning. Phoneme concatenation needs complex computations at the time of execution. Proper phoneme extraction
depends on the context also. A serious problem occurs due to not being able to detect the phoneme boundary accurately.

1.2.3. Diphone concatenation
Diphone is conjunct of two phones that starts at middle of one phone and ends at middle of the next phone. These diphones have no meaning of their own. Generally stable points of phones are preferred to join in concatenated synthesis. Experimentally it is determined that voiced segments of speech signal should be selected for a pleasant output. A problem may arise here due to minimalism of the recording device and speaker variation. The diphones are just collection of some nonsense segments, that is why at the time of concatenation prosody should be imposed on them.

2. Discussion on Subject Topic

2.1. Issues related to concatenative synthesis
Spectral discontinuities and prosodic discontinuities are two major issues in concatenative speech synthesis. Huge mismatching in phonetic and phonemic parameters of the units generally occur in formants of the speech signals at the joining end, ultimately it results in spectral discontinuities. In the same way larger mismatch of pitch values at joining end gives prosodic discontinuities. All these discontinuities may lead to degraded speech output. Natural utterance is obstructed due to lack of proper rhythm and intonation[10]. As mentioned earlier concatenative synthesis builds synthesizer from already stored database of speech units. Unit selection (waveform) method makes choice of appropriate units and concatenate them under some signal processing constraints. Diphone concatenation is another technique, which joins diphones to produce streamline speech. In case of diphone synthesis the phones are associated with duration and F0 target values.

2.2. Syllable as units of concatenation
Unit selection concatenative synthesis is a method where units are selected depending on methodology used (word, diphone or syllable concatenation) and some principles like syllable structure, stress level of neighbouring phones, word level affects etc. In most of the Indian languages, basic linguistic unit is syllable. In case of concatenative speech synthesis syllable plays leading role for natural output by computer. For concatenated synthesis unit size selection largely depends on linguistic characteristic of that language. If the language has smaller number of total syllables then output speech is more natural[11]. Syllable based synthesis has lesser discontinuities and lesser number of joints compared to diphone and phone concatenation. A syllable is generally in the form of V, VC/CV, CCV, CCVC, VCV. Here C means consonant and V means vowel. In most of the cases at least one vowel is needed to form a syllable. For example Assamese word আমি (ami) has two syllables /াম/ and /মি/.

2.3. Introduction to Festival
Platform of this work is Festival. It is an open source software and main purpose of this software is to build speech synthesizer. It has many flexible modules that can be fitted for any language with required modifications. Festival has a number of stand alone tools which can be developed iteratively. All these qualities make festival a greater success. The statistical tools used here can be put together by writing scripts. It uses a separate voice building platform called festvox. Alan W Black and Kevin A Lenzo of Carnegie Mellon University extended this festvox[12] from festival. A number of re writable modules are also integrated with festival platform, such as phrasing, duration modeling, G2P (Grapheme to Phoneme) rules formation etc. Festival is already applied for synthesis of many different languages in the world.

2.4. Different features of Festival
The main data structure used in festival is relational table. Complicated relations are transformed into tree. Overlapping relations are also used here. This overlapping facility saves memory spaces and reduce redundancy. The explained approach is based on unit selection speech synthesis and syllable is
used as basic unit. It is a rule based approach where letter to sound or grapheme to phoneme rules are added for appropriate pronunciation. Festival uses an Utterance data structure where pronunciation of the each speech unit and their prosodic information are found. Prosodic information means information on pitch, duration, stress etc. for individual and its adjacent units. Outputs are stored in tabular format. Sometimes one item may belong to more than one relational tables. Here word item means phrase, phone or word. Every item is assortment of some features or attributes. For example a relational table “phone” is a table where all members are phones with various attributes. At the same time these phones may belong to another relation table “word” also. These types of common features are frequently seen in different tables in festival. The flexible architecture of festival permits execution time description of items as well as relations. So from run time entities also some features can be extracted. Items are generally represented by their identifying name and feature values.

2.5. Methodology

The conversion process of text to speech is mainly divided into two phases; Front end and Back end. Text normalization and all pre processing works are done in the first phase. Assamese is a syllabic centric language and formatted output is achieved after a number of steps. In this paper we are trying to focus on the second phase called “Working on the festival framework.” A few points on pre-processing are discussed here.

The input text to be synthesized must be in UTF-8 format. Its transliterated version is required for processing and a scheme program or an inbuilt tool can serve this purpose. The non standard texts like acronym should be expanded, number should be converted to appropriate text format, grapheme to phoneme rules or pronunciation dictionary is consulted to get proper pronunciation of the input text, syllables of the words are extracted using some algorithms or tools. Acoustic pieces of speech are generated by pronunciation generation part. Phone sequences are the input to this phase and output is the speech unit. Another module called prosodic phrasing divides the input text to meaningful sets of information. There are some other models which were used to identify the prosodic phrasing. Next there is segment duration generation module. This module assigns time to every speech unit known as execution time. This module is important to incorporate tune to the output speech. Last but not least intonation generation module generates melody of the speech. This unit generates fundamental frequency (F0) for output voice signal. There are other intonation generation modules like Tobby and Tilt are also available but for festival tilt model is preferred for its compatibility with festival. This model can predict upward and downward motion of speech signal and thus produce the F0 contour for them.
Figure 1 is the generalized block diagram of Text to Speech synthesizer. As shown in the figure input text should be in UTF-8 format for further processing. Collection of text was done from different sources like Assamese novel and story book. Next normalization is done to get accurate text by converting numbers to suitable format, acronym expansion, word segmentation etc. Pre-processing, normalization and linguistic analysis are performed to get phone sequences.

2.5.1. Token identification

In any programming language we can find a number of tokens. In case of speech synthesis token recognition is accomplished in the first phase. There are a number of methods which are used to recognize tokens of the language. A kind of homograph disambiguation occurs when a number is identified as token. Homograph disambiguation occurs due to different pronunciation of the same identifier. For example a number may mean a date, phone number, rupees or a general numeric value depending upon the context. For every one of them the pronunciation will be different. One solution for this problem is by writing a grapheme to phoneme rules. Fortunately Assamese language has very few disambiguation.

2.5.2. Grapheme to phoneme conversion

Grapheme to phoneme rules means a set of production rules that shows how a word should be pronounced. If the language is simple enough then direct pronunciation can be done easily. Here simple means mapping between written form of the script and pronunciation that is almost one to one. Languages like Spanish, Kanada etc. belong to above mentioned category. But for some languages like English, one orthography may mapped to more than one pronunciations. As a result ambiguity occurs and need a rigorous set of G2P rules and pronunciation dictionary. Assamese language also need G2P rules because some of the pronunciations are not one to one. For example Assamese word  (k2la) has two meanings: black color and deaf person. We can get many more examples like this. Already a number of G2P rules are formulated by many researchers for Assamese language.

There are basically three types of approaches for building G2P rules. They are Rule based approach, Data driven approach and Statistical approach. Rule based approach is most suitable for Assamese G2P rules. At first a word is first searched in the dictionary for its pronunciation. If not found then already built G2P rules are tried for matching. This type of dictionary prepartion is an iterative process and can be updated for future reference. With every lexicon, its Parts of Speech (POS) is also stored in the dictionary. It is because pronunciation varies depending on Parts of Speech.
Same word while used as a noun will be pronounced differently than verb. The above example क़ा (kā) is applicable in this context.

2.5.3. Prosodic analysis

Generation of duration and prosody of the speech units are hard to assume as very less information about them are present in the text to be synthesized. Same sentence is uttered differently (Question asking, Exclamatory, Fact telling) in different purposes. Utterance length or durations of the units are also depended on the inherent meaning of the sentence. This problem can be solved with the help of algorithms, machine learning technique or any simple rule. Simplest approach may be assign a fix average duration to each unit. In general 125ms duration is assigned to every normal unit.

The last phase of a concatenative TTS is working on framework. As already discussed, festival works on Assamese language as a speech generator. There are other two systems, “Speech tools” and “festvox”. An input text is synthesized with the help of festival but voice is generated with the help of festvox. Voice can be built with assist of festvox by users own or other’s voice. A set of commands are used to make festvox workable and compatible with festival. To build a system which can speak like a human, lots of conditions should be brought together. Festvox makes all these trouble easier for users.

To work with festival by considering syllable as basic unit to concatenate, some changes have to be done to the existing system in code and command level. These modifications are done according to characteristics and requirement of the specific language. Some of the necessary changes are mentioned below.

1. In festvox a unique directory have been created for unit selection voice set up. The following changes have been done as a new voice set up.
   a. Make some changes to the original voice set up file.
   b. Normally no phone set is defined for language. So a phone set has to be created.
   c. In original festival phones were considered for concatenation, for Assamese language syllable is used to concatenate, so phones should be replaced by syllable and this replacement should be reflected in all the databases. All phoneme labeled files will be changed to syllable label file.
   d. If symbols are present in the tokenizer set then those should be removed.
   e. The most necessary thing is to create a pronunciation dictionary and define a set of grapheme to phoneme rules. One separate programming in scheme can be done for this.

2. After completing the above steps, commands are set for the following works:
   a. The prompts are generated in the 2nd phase. Actually for the text transcription the prompts are generated. After generation of the prompts they are recorded with very distinctive and proper pronunciation preferably done by a native expert speaker.
   b. Next automatic labeling is done. DonLabel tool is generally used for labeling Indian languages, this tool was generated by a team of Indian researchers[13]. Manually also more correct prompt labeling can be done. In festival ehmm algorithm is used for labeling as it comes together with festvox distribution.
   c. In the next step command is given to generate the utterance file from the label file.
   d. Extracting and fixing pitch marks are done in the next phase.
   e. Mel Cepstral coefficients are generated in the next phase.
   f. Utterance structure should be generated after that.
   g. Cluster units are generated which takes a bit countable time, as it is a searching and matching procedure on a huge dataset. At last it generates catalogue and tree files.
   h. In the last phase speech waveform is generated.

Designing the prompt then generating them have to be done with utmost care. The festvox command generates one file for every prompt and stores them in utterance directory. Each file contains the necessary information like syllable, phones, POS, F0(fundamental frequency), syllable stress, duration etc.
Mel cepstrum attributes are useful at the time of clustering the units. Acoustic characteristics are necessary to identify similar units. As mentioned above they are built just after fixing the pitch marks. It is done with the help of a script that produces these parameters. Some definite techniques are adopted at the time of clustering the similar units.

Grouping or clustering of the speech units are done based on some of the information found at the time of synthesis. Some of the non acoustic information such as F0 and duration values of the units, syllable position in the word, stressing, accents, syllabic boundaries etc. are more important. There are some syllabic centric cost measurements in this synthesis and appropriate ones are selected very cautiously. Syllables are grouped to three categories depending on its position in the word, i.e. at beginning, middle and end of a word. When a syllable is at beginning of a word, searching for its pronunciation is done in that cluster where beginning syllable with identical pitch, duration, accent values are found. Otherwise spectrum mismatch occurs at the joining points of the units thereby unpleasant utterance is created. Below a list of actions performed by different instructions for clustering similar units are shown.

- Speech database collection.
- Build the utterance structure taking syllable as basic unit.
- Build acoustic distance coefficient using LPC method.
- Construct distance table after calculating distances between similar units.
- Building cluster trees from the above table with the help of Wagon tool.
- Create the speech output using festvox command.

In 1.3.1 version of festival, clunits module is used for cluster generation. The process is almost automatic. There is already a stored database with necessary utterance structure and attributes. Build clunits function is used to generate cluster trees and distance tables. After a sequence of modifications to the variables used in the functions, speech output is generated for the input text.

3. Conclusion and Future approach

This article highlights some issues related with building a text to speech synthesizer using unit selection concatenative approach. From the entire discussion above in addition to different research studies on Indian languages, unit selection concatenative synthesis is a promising way for building a synthesizer for Assamese language. The approach mentioned here is syllable centric, means the units to be concatenated is syllable. We are concentrating on selection of highly pronounced syllabi, that are of significantly high frequency. High frequency syllabi means the syllabi which occurs most frequently in a text. Computer program can be written to syllabify words of a language. In this approach sentences were selected from article of one of the Assamese news channel with various words in different context. This selection is based on those sentences which carries at least one high frequency syllable.

There are huge number of works coming up on speech synthesis for North East Indian languages. But without thorough knowledge on phonemes, diphthons, syllable of a language no one can improve more. In this way of synthesis clustering technique is used for selection of the units. But from observation it has been found that biclustering or multi dimension clustering instead of individual parameter clustering may be an easy approach. To select the proper units for concatenation fuzzy logic may be used for computing problems and it may be a good new approach. Above all to get natural sound from any concatenative synthesizer we should avoid the following consequences.

- Less recording of speech.
- Less number of runtime syllable clusters.
- Defect at concatenation points.
References

[1] Flanagan J L 1972 *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, second edition, New York

[2] Silverman K Beckman M B Pirelli, Ostendorf J., M., Wightman C., Price P. Pierrehumbert J and Hirschberg J 1992 *Tobi: A standard for labeling English prosody* in Proceedings of Int. Conf. Spoken Language Processing. pp. 867–870

[3] Thomas S 2007 *Natural sounding text-to-speech synthesis Based on syllable-like units* Research thesis, IIT madras

[4] Daniel Jurafsky James H Martin 2019 *Speech and Language Processing* Eleventh Edition, Published by Pearson India Education Services. Pvt. Ltd, ISBN 978-93-325-1841-4

[5] Kakati Banikanta 2000 *Assamese its formation and development* 5th ed Guwahati India LBS publication

[6] Bordoloi, B N et al 1987 *Tribes of Assam Part-I. Tribal Research Institute* Guwahati 28 Assam,

[7] Ronanki S Reddy S Bollepalli B King S 2016 *DNN-based Speech Synthesis for Indian Languages from ASCII text*. arXiv preprint arXiv:1608.05374

[8] Panda S P Nayak A K 2018 Vowel onset point based waveform concatenation technique for intelligible speech synthesis In Proc: International Conference on Computing Methodologies and Communication (ICCMC 2017) IEEE pp. 622–626

[9] Coleman John 2005 *Introducing Speech and Language Processing* Cambridge University Press UK

[10] Nwakanma Ifeanyi & Oluigbo, Ikenna & Izunna, Okpala. 2014 *Text – To – Speech Synthesis (TTS).* International Journal of Research in Information Technology, Vol 2, Issue 5, May

[11] Panda S P Nayak A.K. & Rai S C 2020 A survey on speech synthesis techniques in Indian languages. Multimedia Systems, pp 453–478, https://doi.org/10.1007/s00530-020-00659-4

[12] Black A W and Lenzo K A 2003 *Building synthetic voices* http://festvox.org/bsv/

[13] Patil Hemant A et al. 2013 *A syllable-based framework for unit selection synthesis in 13 Indian languages* 2013 International Conference Oriental COCOSDA (O-COCOSDA/CASLRE). IEEE