Assamese Dialect Translation System- A preliminary proposal

Sanghamitra Nath, Himangshu Sarma and Utpal Sharma
Department of Computer Science and Engineering, Tezpur University, Assam, India
Email: s.nath@tezu.ernet.in, himangshu.tezu@gmail.com, utpal@tezu.ernet.in

Abstract— Assamese, spoken by most natives of the state of Assam in the North East of India, is a language which has originated from the Indo-European family. However the language when spoken in the different regions of Assam is seen to vary, mostly in terms of pronunciation, intonation and vocabulary. This has given birth to several regional dialects. Central Assamese is spoken in the Nagaon district of Assam and in neighboring areas. Eastern Assamese is spoken in and around the district of Sibsagar. Kamrupia or Kamrupi is spoken in the districts of Kamrup, Nalbari, Barpeta, Darrang, Kokrajhar, and Bongaigoan. In addition we have Assamese variants like Nalbaria, Barpetia, Kamrupia, Goalparia, Jorhatia, etc spoken by people of the respective regions. It would be useful therefore if a system is designed to help two people speaking different varieties of Assamese to communicate with each other. In other words, a seamless integration of a speech recognition module, a machine translation module and a speech synthesis module would help facilitate the exchange of information between two people speaking two different dialects. The proposed system for Assamese Dialect translation and synthesis should therefore, work in such a manner that given a sentence in standard Assamese (written form of the language), it should translate and synthesize the utterances in the dialect requested, or vice versa. Or given a sentence in a particular dialect, it should translate and synthesize the utterances in the dialect requested. The outcome of this work is two folds. Firstly, it will help people understand and appreciate the interesting differences in the Assamese dialects, which is important in preserving the culture preserved in the dialects. Secondly, the proposed system will act as an aid to people speaking different dialects while communicating with each other.

Keywords— Assamese, Dialect, Speech Corpus, Prosody, Speech Synthesis

I. INTRODUCTION

Assamese is the principal language of the state of Assam in North East India and is regarded as the lingua-franca of the whole of North East India. The Assamese language is the easternmost member of the Indo- European family and is spoken by most natives of the state of Assam. As reported by RCILTS, IITG, over 15.3 million people speak Assamese as the first language and including those who speak it as a second language, a total of 20 million speak Assamese primarily in the northeastern state of Assam and in some parts of the neighboring states of West Bengal, Meghalaya, Arunachal Pradesh and other North East Indian states[1]. The Assamese language grew out of Sanskrit, however, the original inhabitants of Assam like the bodos and kacharis had a great influence on its vocabulary, phonology and grammar. Assamese and the cognate languages, Maithili, Bengali and Oriya, is believed to have developed from Magadhi Prakrit which is the eastern branch of the Apabhramsa that followed Prakrit. Mayang is the form of Assamese spoken by the somewhat marginalized Mayang tribe in the northern regions of Manipur while Jharwa Assamese is a pidgin language incorporating elements of Assamese, Hindi and English. Although in the middle of the 19th century, the official language of Assam was considered to be Bengali, British colonizers decreed Eastern Assamese to be the standard form of the Assamese language. Presently, however, Central Assamese is accepted as the principal or standard dialect. Several regional dialects are typically recognized. These dialects are seen to vary primarily with respect to phonology and morphology. However a high degree of mutual intelligibility is shared by the dialects. Dr. Banikanta Kakati, an eminent linguist of Assam, divided the Assamese dialects into two major groups, Eastern Assamese and Western Assamese. However, recent studies have shown that there are four major dialect groups listed below from east to west[2]:

I. Eastern group spoken in and other districts around Sibsagar district.

II. Central group spoken in present Nagaon district and adjoining areas.

III. Kamrupi group spoken in undivided Kamrup, Nalbari, Barpeta, Darrang, Kokrajhar and Bongaigaon.

IV. Goalparia group spoken in Goalpara, Dhubri, Kokrajhar and Bongaigaon districts

The script used by the Assamese language is a variant of the Eastern Nagari Script which traces its descent from the Gupta Script. The script is similar to that of Bengali except for the symbols for /r/ and /w/ and highly

[1] http://www.iitg.ernet.in/rcilts/pdf/assamese.pdf
[2] http://en.wikipedia.org/wiki/Assamese_language

resembles the Devanagiri script of Hindi, Sanskrit and other related Indic languages. It is a syllabary script and is written from left to right. The Assamese alphabet consists of 12 vowel graphemes and 52 consonant graphemes[3]. Both phonemes and allophones are represented in this set of alphabets. Assamese spelling is not always phonetically based. Current Assamese spelling practices are based on Sanskrit spelling, as introduced in Hemkosh, the second Assamese dictionary written in the middle of the 19th century by Hemchandra Baruah which in fact is considered as the standard reference of the Assamese language.

II. BACKGROUND

Though much work has been carried out in the field of Speech translation and synthesis, work related to dialect translation and synthesis specially in the Assamese dialects is near to none. As an initial measure, a survey has been made in the areas of pronunciation modeling, speech recognition, translation and other current speech processing approaches which will be required for the proposed system/work.

A. *Building of Speech Corpus*

A spoken language system, whether it is a speech recognition system or a speech synthesis system, always starts with the building of speech corpora [2]. Therefore to begin with, speech samples need to be recorded and recognized to generate a text file. Then a corresponding speech file with the phonetic representation of the recorded speech is generated and stored. The speech file may be processed in different levels viz., paragraphs, sentences, phrases, words, syllables, phones and diphones and these units of processing are referred to as the speech units of the file. Concatenative speech synthesis involves the concatenation of these basic units to synthesize natural sounding speech. The speech units are added with some more relevant information about each unit using acoustic and prosodic parameters either manually or automatically. Acoustic parameters may be linear prediction coefficients, formants, pitch and gain while prosodic parameters may be duration, intensity or pitch. These parameters are modified by stored knowledge sources corresponding to coarticulation, duration and intonation [10]. The coarticulation rules specify the pattern of joining the basic units while the duration rules modify the inherent duration of the basic units based on the linguistic context in which the units occur. The intonation rules specify the overall pitch contour for the utterance, the fall-rise patterns, resetting phenomena and inherent fundamental frequency of vowels. To make the speech more intelligible and natural, appropriate pauses need to be specified between the syntactic units. The resulting database consisting of the speech units along with their associated information is called the speech corpus. The syllable is a good constituent of the prosodic features of a language making it an appropriate speech unit for most Indian languages. Thus selection of an appropriate speech unit and then annotating it with necessary acoustic and prosodic information play a vital role in the development of Speech corpus [2].

Phones or Syllables as the basic unit

Phones as basic speech units reduce the size of the database because the number of phones for most Indian languages is well below 50. But phones provide very less co-articulation information across adjacent units thus failing to model the dynamics of speech sound. They therefore, are not considered to be efficient in speech synthesis. Most Indian languages however are syllable centered, with the pronunciations mainly based on syllables. Intelligible speech synthesis is possible for Indian languages with syllable as the basic unit [2]. Syllable units are larger than phones or diphones and can capture co-articulation information much better than phones or diphones. Also, the number of concatenation points decreases when syllable is used as the basic unit. Syllable boundaries are characterized by regions of low energy, providing more prosodic information. A grapheme (letter or a number of letters that represent a phoneme) in Indian languages is close to a syllable. The general format of an Indian language syllable is C*VC*, where C is a consonant, V is a vowel and C* indicates the presence of 0 or more consonants. A total of about 35 consonants and 18 vowels exist in Indian languages. There are defined set of syllabification rules formed by researchers, to produce computationally reasonable syllables. Such rules are bound to vary with respect to a language or there may be specific rules particular to a language. A rule based grapheme to syllable converter can also be used for syllabification [2].

Polysyllables as the basic unit

Polysyllable units are formed using the monosyllable units already present in the database, therefore the quality of synthesis can be improved without the requirement of a new set of units. Such a system uses a large database consisting of syllables, bisyllables and trisyllables. In the synthesizing process, the first matching trisyllable is selected followed by the bisyllable and monosyllable units, as needed. Selecting the largest possible unit in the database reduces the number of co-articulation points thereby improving the quality of synthesized speech [2].

Clustering the speech units

Speech unit clustering leads to selection of the best candidates for synthesis. An acoustic distance measure is defined to measure the distance between two units of the same phone type. Cluster units of the same type are formed by evaluating various factors concerning prosodic and phonetic context. A decision tree is then built based on questions concerning the phonetic and prosodic aspects of the grapheme. The leaves of the decision tree reflect the list of database units that are best possible candidates. At the time of synthesis, for each target, the appropriate decision tree is used to find the best cluster of candidate units and then a search is made to find the best path through the candidate units [2].

---

[3] http://www.lmp.ucla.edu/Profile.aspx?menu=004&LangID=83

Prosody

The Speech synthesis system exhibits its naturalness and intelligibility when the corpus is annotated with prosodic information. The concept of prosody is the combination of stress pattern, rhythm and intonation in a speech. Prosodic modeling can be said to describe the speaker's emotion be it anger, depression, excitement or happiness. Recent investigations suggest that the identification of the vocal features containing such emotion may help to increase the naturalness of synthesized speech. Intonation is simply a variation of speech while speaking. All languages use pitch, as intonation to convey an emotion, for instance to express happiness or anger, to raise a question etc. Intonation modeling therefore is an important task that affects intelligibility and naturalness of the speech. Prosody modeling is generally subdivided into modeling the following constituents of prosody - phrasing, duration, intonation and intensity. Two major approaches for prosody modeling are the rule based approach and the corpus based approach [2]. In the rule based approach, linguistic experts derive a set of rules to model prosodic variations by observing natural speech. In the corpus based approach, a well-designed speech corpus, annotated with various levels of prosodic information is used. The corpus is analyzed automatically to create prosodic models which are then evaluated on test data. Based on the performance on test data, the models are then improved. Syllables have sufficient duration information and so it improves the quality of synthetic speech when used as a duration model. Thus syllables are identified as the best-suited processing units for Indian language Speech synthesis [2].

B. Pronunciation Modelling

Many languages such as English, French, German, and Mandarin, have a documented way of how words are pronounced. The pronunciation of words can normally be found in a dictionary. The pronunciations are typically described using IPA (International Phonetic Alphabets) symbols. In Assamese, though some studies have been done to describe the standard Assamese pronunciation, works on Assamese dialects are limited. The phonology of a language or dialect can be analyzed through perception tests, acoustic phonetic analysis, or speech processing techniques. Perception test is based on the perception of the listener and requires native listeners to listen to some sample of sounds, which differs only in a speech sound. If the listener can distinguish the speech sound, then the speech sound is a phoneme of the language. On the other hand, acoustic phonetic analysis analyzes the acoustic features of the spoken signal using spectrograms. These approaches however require expert knowledge. Speech-processing tools such as phoneme recognizer and automatic speech recognizer can be also be used to derive the phoneme set.

The dialects (both eastern and western) of Goalpara straddle the Assamese-Bengali language boundaries and display features from both languages. The phonemes of eastern Goalpara dialect approach those of Assamese while those of the western dialects approach those of Bengali. The distinctive velar fricative /x/ present in Assamese is present in the eastern dialect, but absent in the western dialect[4]. The dental and cerebral (retroflex) phonemes present in Bengali are found in the western dialect, but they approach the alveolar sound in the eastern dialect in consonance with Assamese. The aspirated /ch/ is present both in Bengali and the western dialect, but is not found in the eastern Goalparia dialect and in standard Assamese.

One of the most prominent features of the Kamrupi dialect group is the use of initial stress, as opposed to the use of penultimate stress in the eastern dialects[5], which effectively shortens the word (komora, Eastern dialect; kumra, Kamrupi dialect). In standard Assamese if a word has two /a/ sounds side-by-side, the first /a/ turns into an / □ / or / □ /, a feature that became prominent in the early Assamese period. In Kamrupi, two consecutive /a/ are tolerated (star: / ta□a / (Kamrupi), / t□□a / (Standard)). Epenthesis Epenthesis or the insertion of an extra sound in a word, is a distinguishing feature in western assamese dialects of Kamrup and Goalpara. Example : kaija and kajia. Epenthetic vowels are the rule in Kamrupi dialects, with even diphthongs and triphthongs appearing in initial syllables (haula in Kamrupi; haluwa in Standard) (kewla in Kamrupi; kewaliya in Standard), and a complete absence of diphthongs in the final syllables.

Besides differences in pronunciation, the vocabulary might also be different. For example, the English name 'guava' in Central Assamese is 'modhuriaam' and in Kamrupi it is 'xofram', 'beautiful' in English is 'dhuniya' in Central Assamese and 'thouga' in Kamrupi. There are words which may not exist in the written form. In terms of grammar, they are similar.

C. Speech Recognition and Machine Translation

In terms of machine translation, there are many approaches for translating speech from one language to another. Rule-Based Machine Translation or RBMT[6], also known as "Knowledge-Based Machine Translation" or the "Classical Approach" of MT, is a general term that denotes machine translation systems based on linguistic information about source and target languages basically retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively. An RBMT system generates output sentences (in some target language) from input sentences (in some source language), on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages involved in a concrete translation task. Current state of the art approaches are statistical machine translation and example based translation. Statistical machine translation is a class of approaches that make use of a combination of probabilistic models to choose the most probable translation, for a sequence of spoken words in the source language, given the target language. On the other hand, example based machine translation systems make use of linguistic knowledge and the main idea behind example based machine translation is translation by analogy. However, there are also approaches that

[4] http://en.wikipedia.org/wiki/Goalpariya_dialect
[5] http://en.wikipedia.org/wiki/Kamrupi_dialect
[6] http://en.wikipedia.org/wiki/Rule-based_machine_translation

include statistical knowledge into example-based machine translation to make the translation more accurate.

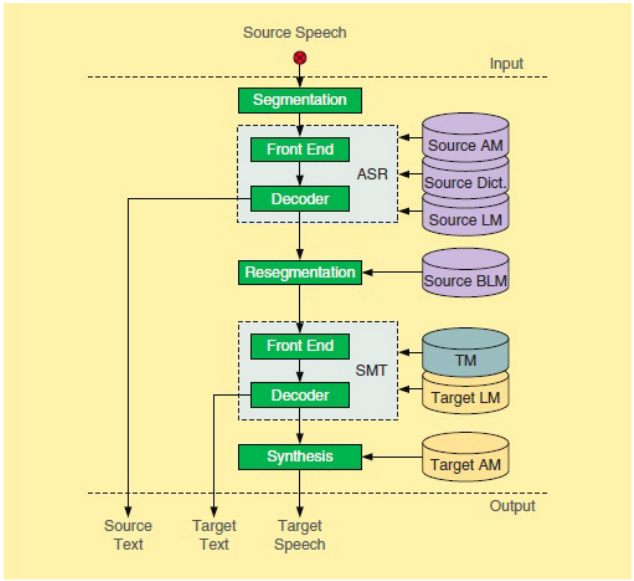

Figure 1: Basic S2S translation chain with a statistical ASR module and an SMT module at its core. *Source: Research Opportunities in Automatic Speech-To-Speech Translation, author: Stuker, S and Herrmann, Teresa and Kolss, Muntsin and Niehues, Jan and Wolfel*

The basic layout of an automatic speech-to-speech translation system presented in Figure 1 s a chain of modules that pass their output to the next module in the chain [5]. The first module performs a segmentation of the audio input. The resulting segments are then transcribed by the automatic speech recognition component. The resulting transcriptions are further processed by a resegmentation, which combines and cuts the speech recognition result into segments that are then translated by the machine translation component. Finally, a speech synthesis module speaks the resulting translation. The said system depicts a setup using statistical machine translation and shows the different components that the individual modules use for their task. While segmentation, resegmentation, and speech synthesis can be seen as supporting techniques, automatic speech recognition and machine translation are fundamental to the operation.

D. Speech Synthesis

Speech synthesis is the artificial production of human speech. With the linguistic knowledge, speech synthesis systems will make use of the information for generating speech given a text. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Speech systems differ in the size of the stored speech units.

Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. The back-end or the synthesizer converts the symbolic linguistic representation into sound. In certain systems, the back-end also includes the computation of the target prosody (pitch contour, phoneme durations), which is then imposed on the output speech.

There are three main approaches to speech synthesis: formant synthesis, articulatory synthesis, and concatenative synthesis [7]. Formant synthesis models the frequencies of speech signal. Articulatory synthesis generates speech by direct modeling of human articulatory behavior. On the other hand, concatenative speech synthesis produces speech by concatenating small, prerecorded units of speech, such as phonemes, diphones and triphones to construct the utterance. A large inventory of speech units of varying prosody is used by most state-of-the-art speech synthesis systems for generating natural sounding speech.

Formant Synthesis

In formant synthesis, the basic assumption is that the vocal tract transfer function can be satisfactorily modeled by simulating formant frequencies and formant amplitudes. The synthesis thus consists of the artificial reconstruction of the formant characteristics to be produced. This is done by exciting a set of resonators by a voicing source or noise generator to achieve the desired speech spectrum, and by controlling the excitation source to simulate either voicing or voicelessness. The addition of a set of anti-resonators furthermore allows the simulation of nasal tract effects, fricatives and plosives. The specification of about 20 or more such parameters can lead to a satisfactory restitution of the speech signal. The advantage of this technique is that its parameters are highly correlated with the production and propagation of sound in the oral tract. The main current drawback of this approach is that automatic techniques of specifying formant parameters are still largely unsatisfactory, and that consequently, the majority of parameters must still be manually optimized.

Articulatory Synthesis

Articulatory synthesis supposedly, is the most satisfying method to produce high quality speech as it generates speech by directly modeling human articulatory behavior. In practice however, it is one of the most difficult methods to implement. The articulatory control parameters include lip aperture, lip protrusion, tongue tip position, tongue tip height, tongue position and tongue height. There are two difficulties in articulatory synthesis. The first difficulty is acquiring data for articulatory model. This data is usually derived from Xray photography. X-ray data do not characterize the masses or degrees of freedom of the articulators. The second difficulty is to find a balance between a highly accurate model and a model that is easy to design and control. In general, the results of articulatory synthesis are not as good as the results of formant synthesis or the results of concatenative synthesis.

Concatenative Synthesis

Concatenative speech synthesis uses phones, diphones, syllables, words or sentences as basic speech units. Speech is synthesized by selecting and concatenating appropriate units

from the speech database or corpus. Large speech units such as words, phrases or sentences improve the quality of synthesized speech however the domain of synthesis is not unrestricted text. When small units such as phones are used, a wide range of words or sentences can be synthesized but with poor speech quality. The most widely used unit in concatenative synthesis is a diphone which is a unit that starts at the middle of one phone and extends to the middle of the following one. Diphones have the advantage of modeling coarticulation by including the transition to the next phone inside the diphone itself. Another widely used unit is the syllable.

Unit Selection Synthesis

In concatenative synthesis, diphones must be modified by signal processing methods to produce the desired prosody. This modification results in artifacts in the speech that can make the speech sound unnatural or robotic. Unit selection synthesis (also, called corpus-based concatenative synthesis) solves this problem by storing multiple instances of each unit with varying prosodies in the database. The unit that matches closest to the target prosody is selected and concatenated so that prosodic modifications needed on the selected unit is either minimized or not necessary at all. Main issues associated with such a system are [3]:

I. Choice of unit size

II. Generation of Speech database

III. Criteria for selection of speech unit

It has been observed that the syllable unit performs better than diphone, phone and half phone, and seems to be a better representation for Indian languages [3]. One limitation of this approach is that the size of the database required is very large, also building such a database is very time consuming since multiple instances of each phone in different contexts need to be stored in the database.

HNM Synthesis

Harmonic plus Noise Models are parametric models which assume that the speech signal is composed of two parts, a harmonic part which responds to the quasiperiodic components of speech and a noise part that responds to the non-periodic components (e.g., fricative or aspiration noise, period-to-period variations of the glottal excitation etc.). The two components are separated in the frequency domain by a time-varying parameter called maximum voiced frequency . The lower band of the spectrum, i.e., upto the maximum voiced frequency is represented by harmonic sinusoids while the upper band is represented by a modulated noise component. Unvoiced parts of speech are represented only by the noise part. Formal listening tests have shown that HNM provides high-quality speech synthesis while outperforming other models for synthesis (e.g., TD-PSOLA) in intelligibility, naturalness, and pleasantness [6].

HMM Synthesis

HMM synthesis ia a statistical parametric speech synthesis system based on hidden Markov models (HMMs) which has grown in popularity over the last few years. This system models spectrum, excitation, and duration of speech simultaneously using context-dependent HMMs and generates speech waveforms from the HMMs themselves. Such a system offers the ability to model different styles without requiring the recording of very large databases [11]. which is highly beneficial in the case of dialects where large databases are not available. HMM-based speech synthesis systems possess several advantages over concatenative synthesis systems [1]. One advantage is that HMM-based systems adapt easily to speakers not present in the training dataset. Speaker adaptation methods used in the field of HMM-based automatic speech recognition (ASR) are adopted for this task. Also less memory is required to store the parameters of the models than to store the data itself. However, the quality of speech generated is not as good as that generated by the Unit Selection Method but the modeling accuracy can be improved by integrating features from other models, like for example, the HNM model may be integrated with the HTS(HMM based speech synthesis system) to have a better quality of speech output [7].

III. PROPOSED ASSAMESE DIALECT TRANSLATION AND SYNTHESIS SYSTEM

We intend to design an Assamese dialect translation and synthesis system as shown in Figure 2 such that given a sentence in standard variety of Assamese, the system will convert the sentence to another variety, i.e. a dialect (and vice versa), and then generate or synthesize the corresponding speech. Once the Assamese Speech Corpus is build, the proposed system for dialect translation is most likely to consist of the following modules: Speech Recognition Module, Assamese Dialect Translation Module , and Speech Synthesis Module.

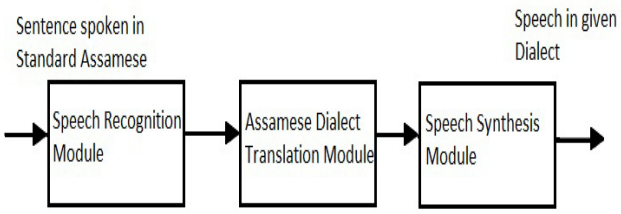

Figure 2: Proposed Assamese Dialect Translation and Synthesis System.

Building of speech corpus

The building of the Speech Corpus starts with text files which are actually lists of words which are commonly used in the standard variety of Assamese and are likely to vary in a particular dialect (say, the Nalbaria variety spoken in and around the district of Nalbari in Assam). Sentences need to be formulated containing commonly used words from the compiled word list. Two groups of speakers are to be

identified, one group consisting of speakers speaking the Nalbaria variety, who have had been born and brought up in Nalbari and had their education in Nalbari. And the other group consisting of speakers of the standard variety of Assamese. The All India Radio speaking style is considered as the standard variety for the proposed system. With the help of these speakers the corresponding speech files need to be recorded in a minimal noise environment. Initially phonetic transcriptions as well as syllabification of these speech files will be carried out manually though automatic segmentation methods will be incorporated later on. Phonetic transcription (or phonetic notation) is the visual representation of speech sounds (or phones). The most common type of phonetic transcription uses a phonetic alphabet, e.g., the International Phonetic Alphabet. Phonetic transcription allows us to step outside of orthography and examine differences in pronunciation between dialects within a given language, as well as to identify changes in pronunciation that may take place over time. For our Assamese Dialect Speech Corpus, the speech unit may be a phoneme, a syllable and in some cases where there is a change in vocabulary, a word.

Speech Recognition Module

The system (initial version) that we propose to design will have an Isolated Word Speech Recognizer as the first module [4]. For each word of the vocabulary, we would want to design separate N-state HMM, where N may be the number of phonemes, syllables or words in a group. Each state may be a phoneme or syllable or we may have a separate HMM for a word. The first task is to build individual word models and use training data to train them. Once the set of HMMs has been designed, trained and optimized, recognition of an unknown word will be performed by setting a score to each word model based upon a set of test observation sequence, and selecting the word whose model score is the highest.

Dialect Translation System

Since the amount of dialectal resources for Assamese is near to none, statistical machine translation is not a good choice. Thus, Example Based Machine Translation (EBMT) or Rule Based Machine Translation (RBMT) approach will be a more reasonable solution. Furthermore, most of the Assamese dialects are similar in terms of grammatical structure; EBMT/RBMT is a more reasonable solution. Once the spoken word is recognized, it may be used to retrieve information from the speech corpus and a RBMT may be done to replace the word with the specified dialect equivalent.

Speech Synthesis System

For the speech synthesis system, HMM speech synthesis may be used for converting text to speech. In an HMM based speech synthesis system [9], the frequency spectrum (vocal tract), fundamental frequency (vocal source), and duration (prosody) of speech are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion. The reasons for selecting HMM speech synthesis system is because it requires less resource, adapts easily to new speakers using relatively small amounts of training data and different transformations can be applied to the HMM acoustic model created. In our case, we want to adapt a speech dialect to a different dialect or may be to the standard language. Once the corpus is ready an acoustic model will be prepared using this corpus along with a pronunciation dictionary. This acoustic model will be used to force align the speaker dependent dialectal speech to an automatic speech recognizer in order to align the phones in the utterances and then train a speaker dependent HMM acoustic model which is phone based. Therefore, given a sentence, the sentence will be first converted to pronunciations before the vocoder can synthesize the corresponding speech [8].

IV. PRELIMINARY WORK

Currently the work is at a preliminary level, where speech corpora are being developed from two speakers speaking the All India Radio variety(AIR) and two speakers speaking the Nalbaria variety of Assamese, taking into consideration all necessary details. The AIR variety which is the form of Assamese spoken by the AIR readers of Assamese news has been considered as the standard form of Assamese and the Nalbaria variety, spoken by the people in and around the district of Nalbari, Assam, is chosen as the dialectal variant. The recordings are being carried out simultaneously at the Jyoti Chitraban Film & Television Institute, Guwahati and Department of CSE, Tezpur University, at a sampling rate of 44.1 kHz and 16 bit resolution. At the same time, a study is being carried out on the speech data collected for both the varieties to find out the basic differences between the two at the segmental, suprasegmental and the subsegmental levels. A preliminary analysis on the recorded data indicates that the vocabulary of the Nalbaria variety can be subcategorized into different groups such as one group where vowel elimination ('mekuri' meaning 'cat' becomes 'mekri' in the Nalbaria variety) in the standard variety leads to the Nalbaria variety, and other groups where a vowel change, vowel elimination with vowel change, change of vowel position, phoneme reduction or a totally new word changes one variety to another. However it is seen that there also exists a group consisting of words which are similar in all respects except for prosodic features such as duration, intensity and pitch. These are the words which though similar phonetically, sound different when spoken by a speaker of the Nalbaria dialect and by a speaker of standard Assamese.

In order to study the role of prosody in dialectal speech and also to find out how each of the prosodic features may be varied to generate dialect speech from the speech of the standard language, an experiment on Prosody Cloning [12] was carried out. Two speakers, prosody donor (speaking Nalbaria) and prosody recipient (speaking standard Assamese) were selected and speech data was recorded. The experiment aimed to find out whether Nalbaria speech can be generated from the standard speech data by selecting different prosodic features of the dialect and imposing them on the standard language. If this is possible, we would also want to find out how each of the prosodic features may be manipulated to convert standard Assamese to Nalbaria and vice versa, or simply to synthesize text in a particular dialect. For this, we

are to make a study of the variations in pitch, intensity and segmental durations.

The experiment as proposed by Kyuchul Yoon[12] was carried out in three steps using the PRAAT software tool. In the first step, segment alignment is carried out and the speech segments of the Nalbaria version are aligned with those of the SL (standard Language). In the second step, the fundamental frequency (F0) contour of the dialect version is super-imposed on the SL version. In the third and final step, the intensity contour of the native version is imposed on the SL version. The resultant waveform resembled the waveform of the Nalbaria utterance to some extent and also sounded close to it.

V. CONCLUSION

In the preliminary stage, the phonetic transcription and syllabification is being carried out manually, however in later stages, automatic segmentation techniques will be used to mark the boundaries of phonemes and syllables. Also, initially the input to the system (to be designed) will be isolated words but the final version of the system should be capable of translating and synthesizing sentences. Building of the speech corpus is of utmost importance because that will decide how well the translated word/sentence is synthesized. So proper choice of speech unit and other additional information like duration, intensity intonation etc. that needs to be stored in the corpus, should be decided upon in advance.

The results of the experiment on prosody cloning show that the cloning can be more accurate if we consider a smaller segment, i.e., a phoneme instead of a syllable, and also if we can take care of the subsegmental variations. The annotation of speech data with pitch, intensity and duration, is under process and we are trying to find out if we can detect some regular variations in it. This would help us in manipulating the prosodic features accordingly and in the places required. Furthermore, our corpus is very small and we cannot come to a conclusion unless we increase the size of the corpus and also the number of speakers in both the dialect and the standard language.

ACKNOWLEDGMENT

The work reported above has been supported by the DeitY sponsored project "Development of Prosodically Guided Phonetic Engine for Searching Speech Databases in Indian languages".

REFERENCES

[1] Matthew Gibson and William Byrne. Unsupervised intralingual and cross-lingual speaker adaptation for hmm-based speech synthesis using two-pass decision tree construction. Audio, Speech, and Language Processing, IEEE Transactions on, 19(4):895{904, 2011.

[2] S Kiruthiga and K Krishnamoorthy. Design issues in developing speech corpus for Indian languages survey. In Computer Communication and Informatics (ICCCI), 2012 International Conference on, pages 1-4. IEEE, 2012.

[3] SP Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal. Experiments with unit selection speech databases for Indian languages. In National seminar on Language Technology Tools, India. Hyderabad, 2003.

[4] Rabiner Lawrence. Fundamentals of speech recognition. Pearson Education India, 2008.

[5] S Stuker, Teresa Herrmann, Muntsin Kolss, Jan Niehues, and M Wolfel. Research opportunities in automatic speech-to-speech translation. Potentials. IEEE, 31(3):26-33, 2012.

[6] Yannis Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. Speech and Audio Processing, IEEE Transactions on, 9(1):21-29, 2001.

[7] Youcef Tabet and Mohamed Boughazi. Speech synthesis techniques. a survey. In Systems, Signal Processing and their Applications (WOSSPA), 2011 7th International Workshop on, pages 67-70. IEEE, 2011.

[8] Tien-Ping Tan, Sang-Seong Goh, and Yen-Min Khaw. A malay dialect translation and synthesis system: Proposal and preliminary system. In Asian Language Processing (IALP), 2012 International Conference on, pages 109-112. IEEE, 2012

[9] Keiichi Tokuda and Heiga Zen. Fundamentals and recent advances in hmm-based speech synthesis. Tutorial of INTERSPEECH, 2009.

[10] B Yegnanarayana, S Rajendran, VR Ramachandran, and AS Madhukumar. Significance of knowledge sources for a text-tospeech system for indian languages. Sadhana, 19(1):147-169, 1994.

[11] Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W Black, and Keiichi Tokuda. The hmm-based speech synthesis system (hts) version 2.0. In Proc. of Sixth ISCA Workshop on Speech Synthesis, pages 294-299, 2007.

[12] Yoon, Kyuchul. "Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody." Journal of the Modern British & American Language & Literature 25.4 (2007): 197-215.