International Journal of Computer Applications (0975 - 8887)
Volume 180 - No.19, February 2018

Incorporating Dialectal Features in Synthesized Speech using Voice Conversion Techniques

Nath Sanghamitra
Assistant Professor
Department of Computer Science and Engineering
Tezpur University

Sharma Utpal
Professor
Department of Computer Science and Engineering
Tezpur University

ABSTRACT

The paper explores to what extent Voice Conversion techniques can help incorporate dialect specific features into synthesized speech. A popular Voice Conversion technique using Gaussian Mixture Models, has been used to develop mapping functions, between speech synthesized by a Text-to-Speech System for the standard form of the language to parallel speech recorded from a speaker of the target dialect. Mel Cepstral Coefficients are used to represent the spectral envelope and pitch, intensity and duration values have been selected to represent the prosody of speech.

General Terms

Dialect Synthesis, Voice Conversion, Gaussian Mixture Models, Text-to-Speech System.

Keywords

Voice Conversion, Gaussian mixture models, Mel Cepstral Coefficients, Formants, F0, Assamese, Nalbaria, Dialect, Pitch, Intensity, Duration, Text-to-Speech System

1. INTRODUCTION

Voice Conversion or popularly known as VC, is the technique of converting speech from a source speaker to speech of a target speaker, and has been used in a number of applications such as conversion of whispered to normal speech, conversion of speaking styles, conversion of emotion, accents etc. VC techniques aim at transforming the characteristics of a speech signal uttered by a source speaker in such a way that the transformed speech sounds like that of the target speaker. Such a conversion transforms not only the organic properties of speech such as voice quality but also linguistic cues such as regional accents and requires transformation of both spectral and prosodic features.

We decided to explore the effects of using VC techniques for incorporating dialectal features into speech synthesized from a Text-to-Speech System (TTS) built for the standard variety of the language. Building a TTS for the standard variety of a language is a much simpler task than building a TTS for a dialect, the main reason being the ready availability of speech data of the standard variety. However if this TTS (for the standard variety) is used to synthesize text from the dialect, the quality of speech deteriorates mainly because of the differences in rules of pronunciation and syllabification, phoneme inventory and prosodic factors such as pitch and duration, between the two varieties of the language. Therefore building a module for post processing the synthesized speech would help in achieving more natural sounding speech. We propose to use VC techniques to bring the speech generated from the standard TTS closer to the target dialect. Our system includes two modules in addition to the TTS module; the preprocessing module which will be a text-to-text translator for translating the text in the standard variety, i.e., All India Radio (AIR) variety of Assamese, to the target dialect Nalbaria. The output of this module, i.e., given an utterance in AIR, the equivalent text in Nalbaria, will be fed to the TTS module. The synthesized utterance will resemble the speech of a non-Nalbaria speaker and will be passed on to the post processing module for incorporating naturalness into the synthesized speech. This module will use VC techniques for converting spectral and prosodic features from the source (speech in AIR) to the target (speech in Nalbaria). The final output will be much closer to the target dialect. The current work focuses on the post processing module. Figure 1 represents the proposed system.

The paper is organized as follows. Section 2 presents a brief introduction to the Assamese language and its dialects, Section 3 presents a brief review of related works, Section 4 describes the building of the speech corpus, feature selection, the VC framework under the heading Methodology, Section 5 describes the building of the mapping function and in section 6, experimental results are presented. Finally some conclusions are drawn from the findings and based on them plans for future work are proposed in Section 7.

2. A BRIEF INTRODUCTION TO THE ASSAMESE LANGUAGE AND ITS DIALECTS

The language we have considered for our study is Assamese, which is the principal language of the state of Assam in 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. It is reported that a total of 20 million speak Assamese primarily in the north-eastern state of Assam and in some parts of the neighboring states of West Bengal, Meghalaya and Arunachal Pradesh and other north-east

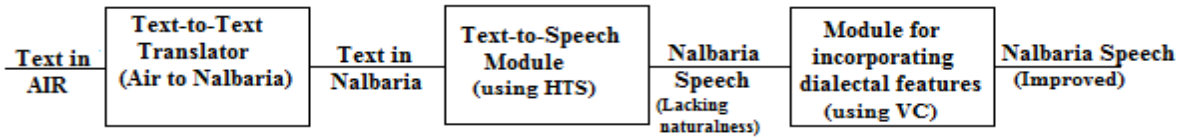

Fig. 1. Block Diagram Representation of Proposed System

Indian states[1]. Several regional dialects are typically recognized. Presently, Central Assamese is accepted as the principal dialect. These dialects vary primarily with respect to phonology and morphology. Recent studies have shown that there are four major dialect groups [2], Eastern group spoken in and other districts around Sibsagar district, Central group spoken in present Nagaon district and adjoining areas, Kamrupi group spoken in undivided Kamrup, Nalbari, Barpeta, Darrang and Goalparia group spoken in Goalpara, Dhubri, Kokrajhar and Bongaigaon districts.

In our study we consider the All India Radio (AIR) variety of Assamese as the standard form of Assamese and the Nalbaria variety (spoken by the people in around the district Nalbari in Assam) which fall under the Kamrupi group as its dialectal variant. The AIR variety is the form of Assamese generally spoken by Assamese news readers of All India Radio. Though a number of dialects of Assamese exist, we have chosen Nalbaria for our preliminary study especially because it greatly differs from the standard form in terms of accent, vocabulary and tempo.

3. LITERATURE REVIEW AND RELATED WORKS

Speech communication is very important in our daily life and it is the most convenient means of communication among humans. A speech signal not only carries linguistic information but also information such as emotion, attitude and speaker's individuality in terms of his gender, age, social status and regional origin. Numerous attempts have been made to improve naturalness of synthesized speech and to make it more human-like. Creation of a new synthesis system is time consuming and expensive since it requires huge amounts of recording and labeling efforts. Therefore researchers are trying to create new voices from existing synthesis systems using various techniques and one such technique is VC which has produced good results.

A dialect may be described as a variation of a given language spoken in a particular place or by a particular group of people, while accents can be considered to originate from variations in articulation habits of a speaker, in his/her own native language. The accent of a speaker may reflect regional affiliation of the speaker therefore the problem of dialect conversion may be reduced to the problem of accent conversion provided the vocabulary and grammar is same. Zheng [1] identifies accent influential acoustic features of two English dialects and investigates accent conversion via formant modification and pitch contour manipulation. Results show that pitch contour modification has a greater effect on accent conversion compared to formant based conversion alone. Zetterholm [2] in her study, mentions that to imitate another speaker's voice and speech behavior, the impersonator has to be aware of not only different markers of group identity such as regional or social dialect, but also personal markers in speech such as pronunciation or articulation.

Accent Conversion may be considered to be a special case of VC where the objective is to capture not the voice quality of the target speaker but the regional accent of the source speaker. VC aims at transforming the characteristics of a speech signal uttered by a source speaker such that the transformed speech sounds like the target speaker. Such a conversion requires transformation of both spectral and prosodic features. Various spectral features like MFCCs, LSPs and LSFs have been used in various works. The most popular VC approach in the literature has been Gaussian mixture model (GMM) based conversion [3]. The data is modeled using a GMM and converted by a function that is a weighted sum of local regression functions. Nonlinear methods using Artificial Neural Networks (ANNs) [4], have been applied to VC to capture the non-linear relationships between input (source) and output (target). Voice quality is known to convey significant variation across different speaking styles and a dialect may also be considered as a style of speech. VC techniques can be used to transform the overall spectral characteristics for realizing corresponding voice quality changes implicitly in the spectral conversion function. Three VC methods, weighted codebook mapping, weighted frame mapping and Joint source-target GMM, used for transforming voice quality of neutral speech to emotional speech have been compared by Turk and Schroder [5].

Regional dialects are known to display differences at the prosodic level too. Prosodic features such as duration, intensity and pitch are commonly used in VC techniques. Srikanth et al. [6] propose a framework for converting both spectral and prosodic features whereby phoneme duration is modified using Gaussian normalized transformation before mapping spectral characteristics of source speaker to target speaker using ANNs. Results confirm that incorporating durational modification has significant improvement over a VC system using only spectral features. Rao et al. [7], present an analysis of duration of sound units with respect to phonological, positional and contextual factors. The most discussed prosodic parameter is pitch or F0, however most VC systems use a linear mean variance method to transform the pitch range of the source speaker to the target speaker overlooking the local variations that affect the speaking style. Other popular methods for pitch conversion are GMMs, Codebook methods [8] and ANNs [9]. Rao et al. [10] use syllable specific features which can capture the inherent relationship between linguistic and production constraints of speech and intensity variation patterns, and feedforward neural networks are used to model syllable intensities.

[1] http://www.iitg.ernet.in/rcilts/pdf/assamese.pdf
[2] http://www.lmp.ucla.edu/Profile.aspx?menu=004&LangID=83

Although a lot of work has been recorded for recognition/identification of dialects, very little work is present for dialect synthesis or bringing naturalness to synthesized dialectal speech. Likewise VC techniques have been used for conversion of speaking styles, accent conversion, emotion conversion etc, but it has not been explored whether such techniques could aid in incorporating dialectal features into synthesized speech. Therefore, the current work is an attempt to apply VC techniques to synthesized speech to make it more natural and also to find out whether conversion of spectral features also contribute towards more natural synthesized speech.

4. METHODOLOGY

4.1 Speech Database

Most VC systems require a parallel database containing the same set of utterances recorded from the source and target speaker. In our case, a set of 50 utterances are recorded from a speaker speaking the target dialect, i.e., Nalbaria, with a Sony recorder in a sound-proof room with a sampling frequency of 16kHz and resolution of 16 bits. This speaker is considered to be our target speaker. The same set of utterances is generated from a TTS built for the standard variety of the Assamese language (AIR) from the transcription of the utterances, with Nalbaria vocabulary, Nalbaria grammar, AIR phonetics, AIR prosodic rules and trained with AIR speech data from the same speaker. In effect, we have a hypothetical person speaking Nalbaria without the knowledge of the phonetics and prosodic rules of Nalbaria and this hypothetical speaker is our source speaker. Since the preprocessing module which is a Text-to-Text translator, is yet to be implemented, we manually transcribe the Nalbaria utterances and feed them to the standard TTS. The transcriptions are carried out by one person and cross-checked by another, both well versed in phonetic transcription and linguistics. However some problems are faced during transcription which need to be reported. These problems can be attributed mainly to the fact that the speakers of Nalbaria (like speakers of most dialects) tend to produce the sounds without full articulatory movement resulting in difficulty of perception. All the four problems are supported by results of analysis of the duration and formant structure of vowels and diphthongs.

Problem 1: Perception of certain sound segments sometimes becomes difficult.
Reason: Speaking rate of Nalbaria in terms of number of syllables per second is high.

Problem 2: Some vowel sounds like /ɔ/, /o/, /u/ are difficult to differentiate. They also sound different from their counterparts in the standard variety.
Reason: Vowels such as /ɔ/, /o/, /u/, /e/, /ɛ/ occupy different locations in the Vowel Space of the two varieties.

Problem 3: Difficult to detect the presence of a diphthong.
Reason: The duration of the secondary vowel in most diphthongs is much smaller.

Problem 4: Difficulty in detecting the presence of /h/ and /r/ sounds before consonants.
Reason: Incomplete articulatory movement.

Mel-Cepstral coefficients (MCEPs) of the order of 21 and pitch values are extracted from both the source utterances (TTS generated) and the target utterances (recorded) using the popular Speech Processing Toolkit (SPTK) after aligning each pair of utterances using Dynamic Time Warping (DTW). MCEPS are extracted using Hanning window, with a frame size of 25ms and a frame period of 5ms. Pitch values are extracted using the RAPT algorithm, with a frame period of 5ms with the upper and lower limit of F0 defined. Furthermore, values of duration are extracted for the vowels, and intensity for the syllables, using the PRAAT[3] tool for phonetic analysis.

4.2 Selection of Features for Conversion

According to the source-filter theory of Speech production, the speech signal can be considered as the output from a linear system, which consists of a source of excitation convolved with the impulse response of a filter [11]. The filter represents the acoustic effects of the vocal tract, which depends not only on the shape and size of the vocal tract but also on the positions of the articulators corresponding to the uttered sounds. MCEPs take human perception sensitivity with respect to frequencies into consideration, and therefore have been selected to represent the filter parameters while fundamental frequency estimates are selected as source parameters. Furthermore, an analysis aimed at studying durational differences between the two varieties is carried out where vowel duration with vowels in word initial, mid and end positions in both varieties, are measured and compared. Results show that mean vowel duration in the Nalbaria variety is smaller compared to that in the AIR variety. Therefore, in addition to MCEPs and pitch, vowel duration can be modified to bring synthesized speech closer to the chosen dialectal variety .

4.3 Building the TTS for the standard variety of Assamese

Initially an existing unit selection TTS built (by IITG) using Festival was used to generate the transcribed Nalbaria utterances. Festival generates the waveform by concatenating appropriate sub-word units from a large database. However since the syllables in the two varieties greatly differ, the syllable-based TTS failed to produce quality output because when a syllable in Nalbaria was not found in the syllable inventory of the standard variety, it was broken into phones and retrieved from the phone inventory of the standard (both varieties have same phoneset), thus introducing discontinuities into the synthesized utterance. Therefore our proposed system needs a Hidden Markov Model (HMM)-based TTS system since such a system is known to simultaneously model spectrum, excitation, and duration of speech using context-dependent HMMs and generate speech waveforms from the HMMs themselves.

In order to build the speech corpus for the TTS, approximately 45 minutes of speech are recorded from a speaker fluent in both the varieties of Assamese, i.e., AIR and Nalbaria variety. This is done mainly to normalise the effects of the vocal tract length which is different for different speakers. The text prompts are prepared with

[3] http://www.fon.hum.uva.nl/praat/

care to include most of the frequently occurring words and also all phones with a minimum frequency of 10. This data is chipped to around 400 wave files and manually transcribed by a person well versed in Nalbaria and then cross checked by another. After preprocessing to remove unwanted pauses, noise, pronunciation errors and clipping, the data is used to build a HMM-based TTS system using the HTS toolkit.

4.4 Voice Conversion Framework

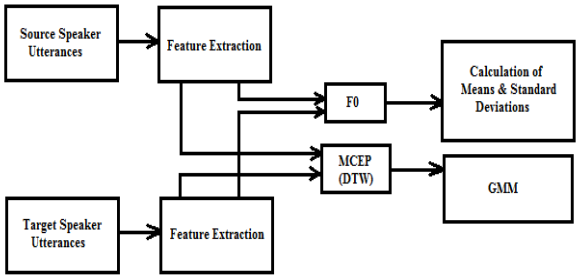

Fig. 2. Training Module in Voice Conversion framework

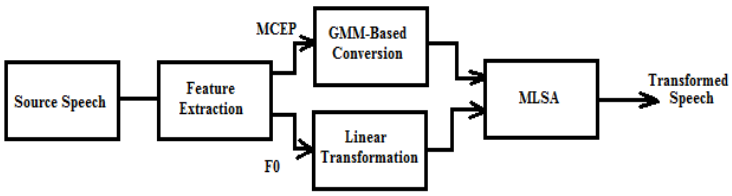

Fig. 3. Testing Module in Voice Conversion framework

Voice conversion is carried out in four basic steps: acoustic modeling, alignment of features, development of a mapping function, and finally synthesis using converted features. During the acoustic modeling step, the short-term spectral properties of the speech signal are captured into a low-dimensional feature vector, for both source and target speech signals. During the alignment step, utterances from the source and the target are time aligned, typically in an automatic fashion by means of DTW or HMMs. During the mapping step, a transformation from source to target features is found through machine learning. The training and testing modules for VC are shown in Figure 2 and Figure 3 respectively. Common mapping techniques include vector quantization [12], ANN [13], and GMM [14], [15].

In our work, we develop mapping functions for the transformation of MCEPs using GMM. At the same time, we transform the pitch contour using the mean-variance method.

4.4.1 Transformation of Pitch and Vowel duration using Linear Mean-Variance method. Since the traditional and most common method of F0 transformation is the Gaussian normalized transformation method or the mean-variance method, we decided to use this method to transform the source speaker F0 to target speaker F0 as indicated in equation 1 .

$$f0_{conv} = \mu_{tgt} + \sigma_{tgt}/\sigma_{src}(f0_{src} - \mu_{src}) \quad (1)$$

where μ_{src} and σ_{src} are the mean and variance of the fundamental frequencies for the source speaker, μ_{tgt} and σ_{tgt} are the mean and variance of the fundamental frequencies for the target speaker, $f0_{src}$ is the source speaker pitch and $f0_{conv}$ is the converted pitch frequency for the target speaker. Before calculating the means and variances of source speech data, the F0 contours of source utterances are corrected manually by comparing with their corresponding target contours. F0 contours for a test utterance and target utterance as well as the converted contour versus target contour are presented in Figure 4 and Figure 5 respectively.

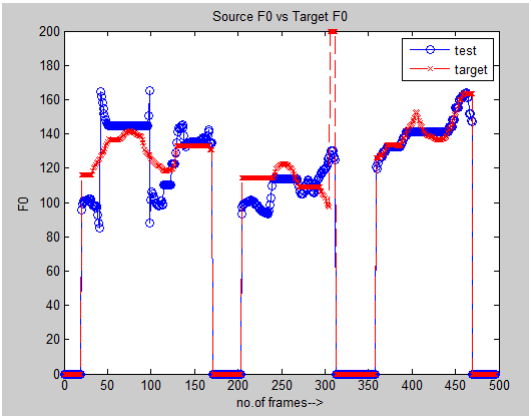

Fig. 4. Source F0 contour vs Target F0 contour

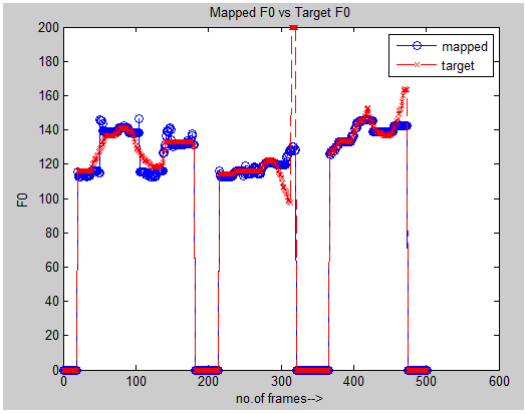

Fig. 5. Mapped F0 contour vs Target F0 contour

Duration of vowels and diphthongs in both source and target speech utterances are measured using a PRAAT script and mean and variance of each of the vowels and diphthongs in source and target data are calculated. The durations of vowels and diphthongs in the test utterance are measured and the above mentioned mean-variance formula is used to transform the durations to match that of the corresponding target. Results of such a conversion for

a test utterance *"tur b^huk nalgiliu k^haba lagbo"* is presented in Figure 6.

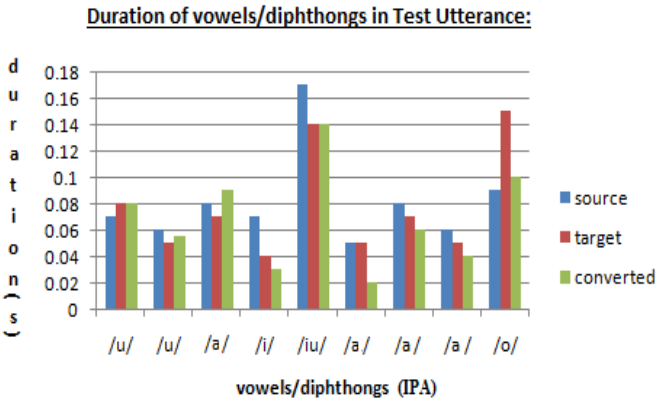

Fig. 6. Transforming vowels/diphthong durations in a Test Utterance

4.4.2 Transformation of Spectral features using GMM. In a GMM based transformation, the learning procedure aims to fit a GMM model to the augmented source and target feature vectors. During the training phase, GMM is adopted to model the distribution of the paired feature sequence z_t, which represents the joint feature vector of source speech vector x_t and target speech vector y_t at frame t. The joint probability density is given as follows:

$$P(z_t|\lambda^{(z)}) = \sum_{m=1}^{M} w_m N(z_t; \mu_m^{(z)}, \Sigma_m^{(z)}) \tag{2}$$

The total number of mixture components is M and w_m is the weight of the m^{th} mixture component. $\lambda^{(z)}$ represents the GMM parameter set consisting of weights, means and covariance matrices for individual mixture components. $\mu_m^{(z)}$ and $\Sigma_m^{(z)}$ are the mean vector and covariance matrix of the m^{th} Gaussian component $N(z_t; \mu_m^{(z)}, \Sigma_m^{(z)})$ respectively and can be expressed in terms of mean vectors of m^{th} mixture component of source and target vectors $\mu_m^{(x)}$, $\mu_m^{(y)}$, covariance matrix of m^{th} mixture component of source and target feature vectors $\Sigma_m^{(xx)}$ and $\Sigma_m^{(yy)}$, cross-covariance matrix of m^{th} mixture component of source and target feature vectors $\Sigma_m^{(xy)}$ and $\Sigma_m^{(yx)}$, in the following manner:

$$\mu_m^{(z)} = \begin{vmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{vmatrix}, \tag{3}$$

$$\Sigma_m^{(z)} = \begin{vmatrix} \Sigma_m^{(xx)} & \Sigma_m^{(xy)} \\ \Sigma_m^{(yx)} & \Sigma_m^{(yy)} \end{vmatrix}. \tag{4}$$

The GMM is then trained using the Expectation Maximization algorithm with the joint source and target vectors which have been aligned with DTW to yield highly robust parameters. The conditional probability density of y_t given x_t can be represented as a GMM as follows:

$$P(y_t|x_t, \lambda^{(z)}) = \sum_{m=1}^{M} P(m|x_t, \lambda^{(z)}) P(y_t|x_t, m, \lambda^{(z)}) \tag{5}$$

where

$$P(m|x_t, \lambda^{(z)}) = \frac{w_m N(x_t; \mu_m^{(x)}, \Sigma_m^{(xx)})}{\sum_{n=1}^{M} w_n N(x_t; \mu_n^{(x)}, \Sigma_n^{(xx)})}, \tag{6}$$

$$P(y_t|x_t, m, \lambda^{(z)}) = N(y_t; E_{m,t}^{(y)}, D_m^{(y)}) \tag{7}$$

Mean vector $E_{m,t}^{(y)}$ and the covariance matrix $D_m^{(y)}$ of the m^{th} conditional probability distribution is written as

$$E_{m,t}^{(y)} = \mu_m^{(y)} + \Sigma_m^{(yx)} \Sigma_m^{(xx)^{-1}} (x_t - \mu_m^{(x)}) \tag{8}$$

$$D_m^{(y)} = \Sigma_m^{(yy)} - \Sigma_m^{(yx)} \Sigma_m^{(xx)^{-1}} \Sigma_m^{(xy)} \tag{9}$$

The converted target feature vector $\hat{y}_t$ is given by

$$\hat{y}_t = E[y_t|x_t] = \sum_{m=1}^{M} P(m|x_t, \lambda^{(z)}) E_{m,t}^{(y)} \tag{10}$$

MCEPs of the order of 21 are employed as acoustic features and are extracted with SPTK. The mapping function is evaluated for different Gaussian mixtures, ranging from M=40, 64, 72 and 128. Results of such a conversion for a test utterance are presented in Figure 7 and Figure 8.

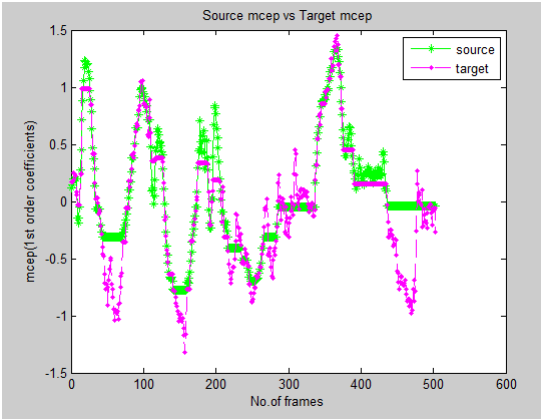

Fig. 7. Source MCEPs vs Target MCEPs of a Test Utterance

4.4.3 Manual Manipulation of Prosodic features. Prosodic features play an important role in bringing naturalness to speech and variation of such features has been observed across dialects in a number of studies. The conversion techniques we have used, converts the spectral features, i.e., MCEPs, F0 global range and vowel/diphthong durations from source to target. In order to bring about the local F0 variations and also the variations in intensity and segmental durations, we have used the PRAAT tool. The

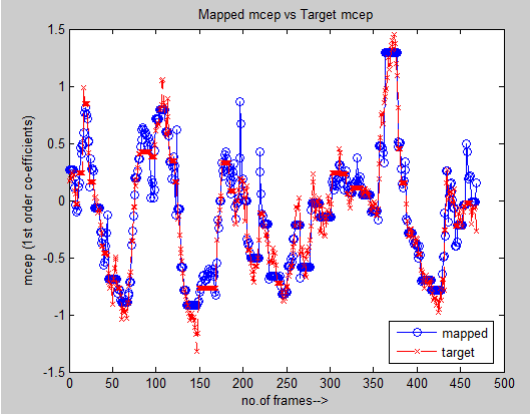

Fig. 8. Converted MCEPs vs Target MCEPs of a Test Utterance

mean intensity of the test utterance is adjusted to match the mean intensity of the target utterance. Secondly, the duration of the vowel and diphthongs in the test utterance are modified to match their counterparts in the target utterance by adding duration points into the duration manipulation object and then modifying the duration points accordingly to lengthen or shorten the segment. Finally the test F0 contour is replaced with the target F0 contour. However this type of a manipulation is possible only when the target utterance corresponding to the test utterance is known.

4.4.4 Evaluation of spectral feature prediction using MCD. Mel Cepstral Distortion (MCD) is an objective error measure known to have correlation with the subjective test results. It is used to measure the quality of voice transformation. MCD is related to filter characteristics and hence is an important measure to check the performance of mapping obtained. MCD is essentially a weighted Euclidean distance defined as follows:

$$MCD = (10/ln10) * sqrt(2 * \sum_{i} (mc_i^t - mc_i^e)^2) \quad (11)$$

where mc_i^t and mc_i^e denote target and estimated MCEPs respectively.
MCD values are calculated between MCEPs of test and target utterances and between corresponding converted MCEPs and MCEPs of target utterance via the GMM-based method.

4.4.5 Evaluation of excitation (pitch), intensity and duration feature prediction using RMSE. The prediction accuracy of the mapping functions used in the VC system for predicting F0 values is evaluated using objective measures such as root mean square error (RMSE). The RMSE is calculated after the durations of predicted contours are normalized with respect to actual contours of target speaker, with the equation below:

$$RMSE = sqrt((\sum_{n=1}^{N} (f0_n^t - f0_n^c)^2)/N) \quad (12)$$

where $f0^t$ and $f0^c$ are the target f0 and converted f0 for each voiced frame and N is the total number of frames per utterance. The f0 conversion is carried out using the mean-variance method.

5. EXPERIMENTAL RESULTS

This section reports the results of various experiments such as GMM based spectral conversion, conversion of F0 and prosody manipulation, carried out on the HTS generated speech data to bring it closer to the Nalbaria variety.

5.1 Comparison of MCD values between test and target MCEPs and converted and target MCEPs

MCD has been adopted as an objective measure to evaluate the conversion of MCEPs from source to target using mapping functions developed by using joint-density GMMs. Table below presents a comparison of MCD scores for 5 test utterances: Since our aim is to find out whether conversion of MCEPs leads to a more natural dialectal utterance, we select the MCEP conversions with the least MCD values for resynthesis.

Table 1. Comparison of MCD values between test and target MCEPs and converted (via GMM method) and target MCEPS

Utterance	mcd_st	mcd_mt (M=40)	mcd_mt (M=64)	mcd_mt (M=72)	mcd_mt (M=128)
Test 1	11.7	8.5	8.6	8.1	7.3
Test 2	11.8	8.5	7.7	7.4	6.0
Test 3	16.5	11.9	11.5	12.0	9.9
Test 4	18.4	13.2	12.2	12.7	10.8
Test 5	11.8	8.3	7.8	7.9	7.0

Here 'M' is the number of Gaussian mixtures used in the conversion.

5.2 Comparison of RMSE values between test and target F0 and converted and target F0

Conversion of f0 for the test utterances is carried out by the mean-variance method. RMSE values between test and target utterances and corresponding converted and target utterances are presented in the table below:
Here again the best f0 conversions are selected for resynthesis.

Table 2. Comparison of RMSE (F0) values

Utterance	rmse_st	rmse_mt
Test 1	19.3	19.3
Test 2	15.1	15.4
Test 3	29.8	29.5
Test 4	17.3	16.6
Test 5	26.0	25.0

5.3 Subjective evaluation using MOS

A Mean Score Opinion (MOS) test is carried out on the synthesized utterances with and without prosodic modification. A total of 10 sets of utterances are given to each of the 10 evaluators who are well versed in the Nalbaria variety of Assamese. Each set consists of 5 utterances, (i) the HTS generated utterance, (ii) the HTS generated

utterance with prosody modification, (iii) the converted (by the VC system) utterance, (iv) the converted utterance with prosody modification and (v) the target utterance spoken in Nalbaria. Since the conversion and prosody modification is carried out on TTS generated utterances, the final utterances are noisy. Our aim is to find out how close the converted and manipulated utterances are to Nalbaria in terms of naturalness, and therefore the evaluators are asked to listen to each of the 5 utterances in each set and give a score using a 5-point scale in terms of naturalness, based on the question 'Which utterance is closest to Nalbaria?' or 'Which utterance is most likely to be spoken by a person speaking Nalbaria?' The table below presents the MOS scores:

Table 3. MOS Results

Sl.No.	Utterance Type	Score(avg)
i	HTS	2.3
ii	HTS + prosodic mod.	2.9
iii	HTS + spectral conv.	3.2
iv	HTS + spectral conv. + prosodic mod.	4.0
v	Target utterance in Nalbaria	5.0

6. CONCLUSION AND FUTURE WORK

Results indicate that the GMM based mapping function has brought the source MCEPs closer to the target MCEPs. The mean-variance method of F0 conversion converted only the global range and the local variations of target contours could not be achieved. This does not affect normal VC where the aim is to convert speech from source to target speaker as the source speaker's F0 range is converted to that of the target. However, since we are trying to convert speech from one dialect to another and it is known that the prosody of a dialect has a direct impact on the local variations of the F0 contour, therefore the inability to map the local variations result in lack of naturalness in the converted utterances. Results of duration conversion using the mean-variance method are not satisfactory. Better results can be achieved by taking into account contextual information such as position of vowel/diphthong in the word, stress, etc. during the conversion. Results of the subjective test presented in Table 3, show that the lowest score is achieved by the utterances generated by the HMM-based TTS for the standard variety of Assamese followed by the same with prosody modification. This indicates that a TTS for the standard variety is not a good option to generate dialectal speech. The highest score is achieved by the utterances whose MCEPs and F0s have been converted using the GMM-based mapping function with manual modification of pitch, duration of vowels/diphthongs and mean intensity of the whole utterance. Results indicate that both spectral and prosodic features carry paralinguistic information and play a crucial role in bringing the TTS (for the standard variety) synthesized utterance closer to the chosen dialect. There is scope of using VC techniques for generating dialectal speech provided the pitch, intensity and duration values are also converted appropriately. The next phase of work may be directed towards developing efficient methods for capturing the local variations of prosodic features from source to target speech. Another future direction that needs to be addressed is the automatic translation of utterances from standard variety to dialectal variety. As of now the transcriptions of the translated utterances are fed as input to the HTS.

7. ACKNOWLEDGEMENTS

The authors thank Samarjit Barman, Jyoti Prasad Talukdar, Himangshu Sarma, Trideep Baruah and Utpal Sharma for lending their voices to build the corpus and are indebted to Mancha Jyoti Malakar for his help in recording and corpus building. They are also thankful to MHRD, Centre of Excellence on MLBDA, for financial support.

8. REFERENCES

[1] Dang Cong Zheng. Accent conversion via formant-based spectral mapping and pitch contour modification. 2011.

[2] Elisabeth Zetterholm. Same speaker–different voices. a study of one impersonator and some of his different imitations. In *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, pages 70–75, 2006.

[3] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Continuous probabilistic transform for voice conversion. *Speech and Audio Processing, IEEE Transactions on*, 6(2):131–142, 1998.

[4] Srinivas Desai, Alan W Black, B Yegnanarayana, and Kishore Prahallad. Spectral mapping using artificial neural networks for voice conversion. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(5):954–964, 2010.

[5] Oytun Türk and Marc Schröder. A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis. In *INTERSPEECH*, pages 2282–2285, 2008.

[6] Ronanki Srikanth, B Bajibabu, and Kishore Prahallad. Duration modelling in voice conversion using artificial neural networks. In *Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on*, pages 556–559. IEEE, 2012.

[7] Krothapalli S Rao, Shashidhar G Koolagudi, et al. Selection of suitable features for modeling the durations of syllables. *Journal of Software Engineering and Applications*, 3(12):1107, 2010.

[8] Zeynep Inanoglu. Transforming pitch in a voice conversion framework. *St. Edmonds College, University of Cambridge, Tech. Rep*, 2003.

[9] Bajibabu Bollepalli, Jonas Beskow, and Joakim Gustafson. Non-linear pitch modification in voice conversion using artificial neural networks. In *Advances in Nonlinear Speech Processing*, pages 97–103. Springer, 2013.

[10] V Ramu Reddy and K Sreenivasa Rao. Intensity modeling for syllable based text-to-speech synthesis. In *Contemporary Computing*, pages 106–117. Springer, 2012.

[11] Gunnar Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 1971.

[12] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)*, 11(2):71–76, 1990.

[13] M Narendranath, Hema A Murthy, S Rajendran, and B Yegnanarayana. Voice conversion using artificial neural networks. In *Automatic Speaker Recognition, Identification and Verification*, 1994.

[14] Alexander Kain and Michael W Macon. Spectral voice conversion for text-to-speech synthesis. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, pages 285–288. IEEE, 1998.

[15] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8):2222–2235, 2007.