EFFECTIVENESS OF PLP-BASED PHONETIC SEGMENTATION FOR SPEECH SYNTHESIS

Nirmesh J. Shah, Bhavik B. Vachhani, Hardik B. Sailor and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, India

E-mail: {nirmesh_shah, vachhani_bhavikkumar, hardik_sailor, hemant_patil}@daiict.ac.in

ABSTRACT

In this paper, use of Viterbi-based algorithm and spectral transition measure (STM)-based algorithm for the task of speech data labeling is being attempted. In the STM framework, we propose use of several *spectral* features such as recently proposed cochlear filter cepstral coefficients (CFCC), perceptual linear prediction cepstral coefficients (PLPCC) and RelAtive SpecTrAl (RASTA)-based PLPCC in addition to Mel frequency cepstral coefficients (MFCC) for phonetic segmentation task. To evaluate effectiveness of these segmentation algorithms, we require manual accurate phoneme-level labeled data which is *not* available for low resourced languages such as Gujarati (one of the official languages of India). In order to measure effectiveness of various segmentation algorithms, HMM-based speech synthesis system (HTS) for Gujarati has been built. From the *subjective* and *objective* evaluations, it is observed that Viterbi-based and STM with PLPCC-based segmentation algorithms work better than other algorithms.

Index Terms— Hidden Markov Model (HMM), Spectral Transition Measure (STM), PLPCC

1. INTRODUCTION

Over the past few decades, there is growing research interest in *text-to-speech* (TTS) synthesis. In particular, there is a need for large amount of reliably segmented speech data for instance, for improving speech recognition and synthesis performance. To segment the speech signal at appropriate sound units of speech, there are mainly two approaches, *viz.*, *manual vs. automatic* segmentation algorithm. Manual segmentation is time-consuming and tedious process (even though manual labeling may have long-term impact). No two human annotators can identify the phonetic boundaries exactly the same. The process of segmentation can be done at various sound units like phone, syllable, word, etc.

The authors would like to thank Department of Electronics and Information Technology (DeitY), Govt. of India for their kind support to carry out this research work (which was partly supported by two DeitY sponsored projects, viz., Development of prosodically guided phonetic engine for searching speech databases and Text-to-speech synthesis system in Indian languages). They also thank consortium leaders, Prof. B. Yegnanarayana (IIIT Hyderabad), and Prof. Hema A. Murthy (IIT Madras). In addition, they thank Prof. T. Nagarajan (SSNCE, Chennai) and all the participants who took part in subjective evaluation of HTS systems.

In this paper, an attempt is made to segment speech at phone-level because phone is a smallest acoustic unit of pronunciation and in any language, number of phones is fixed. An approach using *10*-dimensional (i.e., *10-D*) Mel frequency cepstral coefficients (MFCC) for every *10 ms* for computing a spectral transition measure (STM) in order to capture the effect of spectral rate of change in time-domain is proposed in [1]. This paper proposes use of spectral features such as recently developed Cochlear Filter Cepstral Coefficients (CFCC), Perceptual Linear Prediction Cepstral Coefficients (PLPCC) and RelAtive SpecTrAl (RASTA)-based PLPCC in addition to MFCC during STM calculation for automatic phone segmentation task [2]. Furthermore, Viterbi-based forced alignment algorithm is also used [3],[4]. These algorithms are applied on Gujarati (one of the official languages of India [5]) speech database.

In order to evaluate various automatic phonetic segmentation algorithms, we require manual accurate phoneme-level speech database which is not available for Gujarati language. Now question is how to measure accuracy of various automatic phonetic segmentation algorithms? Hence, it was decided to build real world application (such as TTS) to evaluate the performance of automatic phonetic segmentation algorithms. For Gujarati, unit-selection system (USS)-based approach has been applied recently at syllable-level [6]. However, to the best of authors' knowledge, phone-based TTS for Gujarati is not yet built.

In this work, we have used statistical parametric synthesis (SPS)-based method for building TTS. In particular, we used Hidden Markov Model (HMM)-based Speech Synthesis System (HTS) for developing TTS in Gujarati. In HTS, HMM for each phoneme has to be prepared. Hence, intelligibility as well as quality of synthetic files will directly depend on accuracy of our phoneme-based label files. We have used *subjective* and *objective* evaluation methods to evaluate performance of various HTS voice.

2. PHONETIC SEGMENTATION ALGORITHM

For automatic phonetic segmentation, various approaches such as *supervised* and *unsupervised* methods are proposed. Supervised methods require extensive training on speech material. The training material needs to be transcribed in

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

terms of the speech sound units determined by the suitable segmentation algorithm. On the other hand, unsupervised methods (such as maximum margin clustering [7], jump function [8], group delay [9]) do not require such training. In addition, unsupervised methods are based on human knowledge and understanding of the nature of speech and are therefore language and speech style- independent. These methods simply use appropriate signal processing algorithms followed by peak picking algorithm to detect the acoustic-phonetic boundaries [10].

2.1. Unsupervised STM-based Phonetic Segmentation

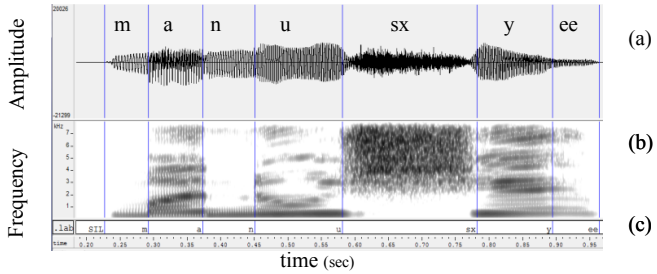

Figure 1: (a) Time-domain speech signal, (b) corresponding phoneme-label, (c) corresponding phoneme-label for an utterance, 'મનુષ્યે', taken from Gujarati database.

From figure 1, it can be seen that spectral variation is very useful to detect phone boundaries. This information is exploited in STM approach [11]. The spectral features used in [1] for speech segmentation is MFCC. In [1], the speech signal is transformed into frames (computed over *30 ms* Hamming window) and then transformed into a set of *10-D* MFCC (excluding the zero-order coefficient that represent the total signal energy) (with computational details given in [12]). The frame rate employed in this study was *100/s* (i.e., *10 ms* frame step or frame increment). After computing MFCC, they used STM to capture the spectral transition between two phones. The STM employed in this study was the same as that of proposed in [11]. This STM, at frame, *i*, can be computed as a mean squared error (MSE) [11], i.e.,

$$C(i) = (\sum_{l=1}^{L} a_l^2(i)) / D, \qquad (1)$$

where *C(i)* is STM at given frame *i*, *D* is the dimension of the spectral feature vector (i.e., *10* in this case) and a_l is the regression coefficient or the rate of change of the spectral feature defined as [11]:

$$a_l(i) = (\sum_{j=-I}^{I} MFCC_l(j+i) * j) / (\sum_{j=-I}^{I} j^2), \qquad (2)$$

where *j* represents the *frame index* and *I* represents the number of frames (on each side of the current frame) used to compute these regression coefficients. It is evident from eq. (2) that STM inherently embeds the delta cepstrum calculations. It is also proposed here to use other spectral features of speech apart from MFCC in STM. For example, Cochlear Filter Cepstral Coefficients (CFCC) is used here which was recently developed Qu Li and Yan Huang primarily for speaker recognition [13]. In addition, perceptual linear predictive cepstral coefficients (PLPCC) [3], [14], [31] and RelAtive SpecTrAl (RASTA) [15] methodology which makes PLP more robust to linear spectral distortion has also been used for computing STM. For computation of these features, *30 ms* frame duration with *10 ms* frame increment is used. In addition, *I=2* is taken for a *10 ms* frame step corresponding to an interval of *40 ms* centered on the current frame at which the *C* value (in eq. (1)) is computed. A larger interval could result in missing some phone boundaries whereas a shorter interval could result in the detection of too many false phone boundaries.

2.1.1. Perceptual Linear Prediction Cepstral Coefficients (PLPCC)

Perceptual linear prediction (PLP) is one of the LP-based analysis methods which incorporate a *nonlinear* frequency scale and other known properties from the psychophysics of hearing [14]. In PLP, Fourier transform is applied to compute the short-term power spectrum, and the perceptual properties are applied while the signal is represented in filterbank. The spectrum is transformed to a Bark scale, and this spectrum is pre-emphasized by a function which approximates the sensitivity of human hearing at different frequencies. The output is compressed to approximate the nonlinear relationship between the *intensity* of a sound and its perceived *loudness*. The all-pole model of LP is then used to give a smooth, compact approximation to the simulated *auditory spectrum*, and finally the LP parameters are transformed using recursive relations to cepstral coefficients for use as segmentation features [3]. PLP analysis is more consistent than conventional LP analysis with human hearing.

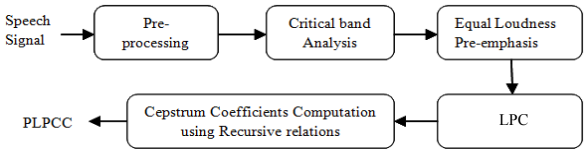

Figure 2:Block diagram for computation of PLPCC. After [3],[30].

The PLP technique (as shown in figure 2) is based on the short-term spectrum of speech. Although the short-term spectrum of speech is subsequently modified by several psychophysically-based spectral transformations, the PLP technique is vulnerable when the short-term spectral values are modified by the frequency response of the communication channel. Human speech perception seems to be less sensitive to such steady-state spectral factors [14].

2.1.2. Labeling of STM-based boundaries

Segmented boundaries are obtained at phone-level using STM algorithm. The issue is to use these segmented boundaries effectively in TTS system development. For that purpose, text material for particular language can be used. This text material is converted into phoneme sequence.

These phones are aligned with automatically estimated boundaries. In this step, the phone or syllable boundaries are estimated automatically.

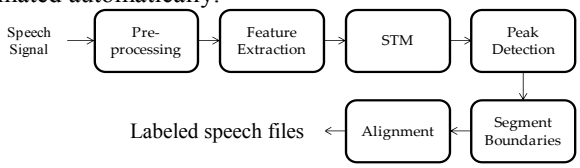

Figure 3: Block diagram for generating segmented labeled file.

For aligning the estimated boundaries with exact number of text, following steps are followed [1].

1. Check whether the number of estimated boundaries (*NE*) from phonetic segmentation algorithm is equal to the number of ground truth boundaries (*NP*) from text material.
2. If not, then check whether it is less and if found, then go to *step 3* or in case it is more, go to *step 6*.
3. If it is less (*NE*<*NP*), then insert *NI*=*NP*-*NE* number of boundaries.
4. For that region, difference between estimated boundaries is taken, and where the difference is larger, the boundaries are inserted. However, the length of that particular segment will depend upon duration of previous label (for this purpose, average phone duration is required).
5. Repeat *step 3*, *step 4*, and *step 5* until *NI*=0.
6. If it is more (*NE*>*NP*), then remove *NR*=*NE*-*NP* number of boundaries.
7. In that case, we merge minimum duration segment to its right or left segment.
8. Repeat *step 6* and *step 7* until *NR*=0.

2.2. Supervised Viterbi-based Phonetic Segmentation

Supervised Viterbi-based alignment algorithm is also used for phonetic segmentation. The process of computing the Viterbi path over training data is called *forced alignment* [4]. Here, we do not know the optimal state sequence assigned to each observation sequence. Following major steps were carried out to generate phoneme-based labeled files [3].

1) Prepare *5* minutes of manual data at phoneme-level.
2) HMM was trained using MFCC features corresponding to each phoneme class [16].
3) Remaining wave files were labeled using speech recognition system with the help of HMM. The only difference between the present work and speech recognition is that phonetic transcription of sentences was already given [3], [16].
4) Now, again HMM was trained using entire speech data.
5) Repeat *step 3* and *step 4* until convergence is achieved.

2.3. Relative performance of various segmentation algorithms

To evaluate relative performance of various segmentation algorithms, *10* minutes of manual phoneme labeled Gujarati speech data was taken. If detected boundary of phoneme is within *10 ms* duration (called as *tolerance interval*) of the manual boundary, then it is considered as correct.

Table 1: Accuracy of various segmentation algorithms

Algorithms	STM MFCC	STM CFCC	STM PLPCC	STM RASTA	Viterbi
Accuracy (%)	40.57	38.60	**40.60**	39.58	**45.86**

Table 1 shows the performance of various segmentation algorithms. From Table 1, it is clear that Viterbi and STM+PLPCC algorithm works on and average better than compared to other algorithms. Creating manual labeled data was very tedious work. Hence, HTS has been built in order to measure relative performance of various segmentation algorithms which is discussed in next section.

3. HMM-BASED SPEECH SYNTHESIS SYSTEM

3.1. Block diagram of HTS

Figure 4 shows the basic block diagram of HTS (which is basically motivated by the study reported in [17]) developed for Gujarati language. The diagram is divided into two parts, *viz.*, *training* part and *synthesis* part.

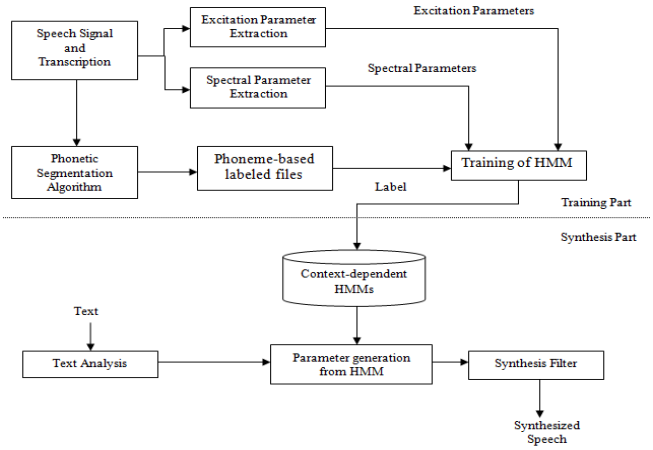

Figure 4: Block diagram of an HTS. After [17].

Training Part: Mel generalized cepstral coefficients (MGC) and their dynamic features are taken as *spectrum* (i.e., vocal tract system) parameters. In addition, *log* F_0 and its dynamic features are taken as *excitation* (i.e., speech source) parameters. With the help of phoneme-based labeled files, context-dependent HMMs for each phoneme for all parameters were prepared and trained. Context-dependent model was selected in order to cover *suprasegmental* information. Thus, spectrum, excitation and durations are going to be modeled in unified framework [18].

Synthesis Part: For given sentence (which has to be synthesized), its corresponding utterance is converted to context-dependent phoneme sequence. Then according to the phoneme sequence, utterance HMM is constructed by concatenating context-dependent HMMs. Then state duration of HMMs is determined from duration model and using speech parameter generation algorithm, spectrum and excitation parameters are generated [19]. Finally, using Mel Log Spectrum Approximation (MLSA) filter speech signal is synthesized [20].

3.2. Adaptation to Gujarati language

In order to develop HTS for Gujarati language, we require Gujarati speech data. We have collected several texts from various sources like news, blogs, stories, etc. We recorded these texts from professional native female and native male voice over artist [21], [6]. We have noticed that except context-dependent modeling, every block diagram of HTS is language-independent, however, contextual information is language-dependent. The HMM framework provides a general setup for sufficient context modeling and which is easily be adapted to other languages [22],[30] For the phonemic representation of Gujarati language, a set of *49* phonemes were taken that are broadly classified into *SIL* (i.e., silence), *35* consonants and *13* vowels. We require different groups of phonemes. For Gujarati language, to do classifications, we have used International Phonetic Alphabet (IPA) chart for Gujarati language for consonants and for vowels [23]. Classifications of phonemes were used for *question se*t preparation.

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

In this paper, *10* different HTS systems each on *3* hours of Gujarati data from male and female voice over artist has been built using various automatic phonetic segmentation algorithms. *105*-dimensional feature vector, *viz.*, Mel generalized cepstral (MGC) coefficients (including 0^{th}, delta and delta-delta coefficients) per frame (of durations *25 ms* and frame shift of *5 ms*) and *3-D* $\log F_0$ per frame has been taken. *Penta-phone* level context-dependent modeling has been used [29]. Since HTS does not store speech sound units rather it stores statistical parameters, size of HTS developed for Gujarati is within range of *3 MB* to *3.77 MB*.

4.2. Subjective Evaluation

Five synthesized utterances (from held out database) from each HTS system was taken and mean opinion score (MOS) [24] analysis and degraded MOS [25] (i.e., DMOS) from *20* Gujarati native subjects. Files from different systems are played randomly and subjects were asked to give score in *1* to *5*. Then from these scores, MOS and DMOS are calculated for each system.

Table 2: MOS, DMOS and WER analysis of developed HTS

Algorithm	DMOS		MOS		WER (%)	
	M	F	M	F	M	F
Viterbi	**2.82**	**3.04**	**2.65**	**2.85**	**2.68**	7.26
STM+PLPCC	**2.71**	**2.72**	**2.55**	**2.55**	**3.96**	**3.7**
STM+MFCC	2.50	**2.72**	2.00	2.00	11.9	21.9
STM+CFCC	2.13	2.17	2.35	2.5	4.98	10.2
STM+RASTA	2.39	2.04	2.25	1.88	10.7	13.6

Intelligibility test was performed in which they have to transcribe synthesized utterances and based on that word error rate (%) was calculated by following formula [26],

$$WER(\%) = \left(\frac{I + D + S}{T} \right) \times 100, \qquad (3)$$

where I is the total number of *insertion*, D is total number of *deletion* and S is the total number of *substitution* that one needs to do in order to make transcription *100 %* correct at word-level and T is the total number of words. From Table 2, it is clear that Viterbi-based and STM with PLPCC features perform better compared to other algorithms. In addition, even though MOS is not that much impressive of our Gujarati systems, however, it is found to be highly intelligible that is evident from WER analysis.

4.3. Objective Evaluation

Widely used *Mel Cepstral Distance* (MCD) measure is used for measuring the accuracy of the spectral envelope of the synthetic speech with respect to natural speech [26]. We have used *Dynamic Time Warping* (DTW)-based approach to align natural speech and synthetic speech as both have different durations [27]. Normalized DTW-based Mel cepstral distance is given by:

$$MCD \triangleq \frac{Distance\ calculated\ from\ DTW}{No.of\ points\ during\ back\ traced}, \qquad (4)$$

where number of points during back traced in DTW is used as normalizing factor for distance calculated [28].

Table 3: MCD analysis of different HTS systems

Algorithm	Female	Male
Viterbi	**3.750**	**3.421**
STM+PLPCC	**4.005**	3.440
STM+MFCC	4.056	**3.415**
STM+CFCC	4.066	3.444
STM+RASTA	4.059	3.446

From the Table 3, it is evident that in case of female voice, Viterbi-based approach and STM with PLPCC-based system gives less MCD score (i.e., performs better). In case of male voice, all methods of labeling have almost similar results. However, STM with MFCC, PLPCC and Viterbi-based method has slightly less MCD score than other methods. Hence, objective measures almost agree with subjective evaluation which is evident from Table 3.

5. SUMMARY AND CONCLUSIONS

In this work, various automatic phonetic segmentation algorithms have been applied to Gujarati language. In addition, to evaluate their relative performance, different HTS systems have been built on same amount of training data where label files are generated using different automatic phonetic segmentation algorithms. Based on subjective and objective evaluations, we found that Viterbi-based and STM with PLPCC features-based HTS systems work on and average better than other HTS systems. One of the possible reasons for getting better results with PLPCC is that it uses critical band masking curves and its equal loudness pre-emphases in feature extraction. We are able to achieve very high *intelligibility* with HTS system, however, it is lacking *naturalness*. Hence, our future research efforts will be devoted towards improving performance of TTS in terms of naturalness.

REFERENCES

[1] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," in INTERSPEECH, Pennsylvania, pp. 645-648, 2006.

[2] B. B. Vachhani and H. A. Patil, "Use of PLP cepstral features for phonetic segmentation, " in International conference on Asian Language Processing (IALP), Urumqi, pp. 143-146, 2013.

[3] L. R. Rabiner, "A tutorial on hidden Markov model and selected applications in speech recognition," Proc. of the IEEE, vol. 77, no. 02, pp. 257-286, February 1989.

[4] L. R. Rabiner and B. H. Juang, Fundamentals of speech recognition.: Prentice-Hall,USA, 1993.

[5] [Online]. http ://www.lisindia.net/Gujrathi/Gujarathi.html (Last Accessed October 27th, 2013).

[6] H. A. Patil et. al., "Algorithm for speech segmentation at syllable level for text-to-speech synthesis system in Gujarati," to appear in Oriental International Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA) Conference, Gurgaon, India, 2013.

[7] P. Estevan, O. Scharenborg, and V. Wan, "Finding maximum margin segments in speech ," in International Conference on Accoustics, Speech and Signal Proceesing (ICASSP), Honolulu, pp. 937-940, 2007.

[8] G. Aversano, A. Esposito, and M. Marinaro, "A new text-independent method for phoneme segmentation," Proceedings of the 44th IEEE Midwest Symposium on Circuits and Systems, vol. 2, pp. 516-519, 2001.

[9] L. Golipour and D. O'Shaughnessy, "A new approach for phoneme segmentation of speech signals," in INTERSPEECH, Belgium, pp. 1933-1936, 2007.

[10] O. Scharenborg, V. Wan, and M. Ernestus, "Unsupervised speech segmentation: an analysis of the hypothesized phone boundaries," J. Acoust. Soc. Amer., vol. 127, no. 2, pp. 1084-1095, 2010.

[11] S. Furui, "On the role of spectral transition for speech perception," J. Acoust. Soc. of Amer., vol. 80, no. 4, pp. 1016-1025, 1986.

[12] C. Ling, D. Minghui, and C. Paul, "Segmentation of speech signals in template-based speech to singing conversion," in APSIPA Annual Summit Conference (APSIPA ASC), Xi, An China, pp. 1-4, 2011.

[13] Q. Li and Y. Huang, "An auditory based feature extraction algorithm for robust speaker identification under mismatched conditions ," IEEE Trans. on Audio, Speech and Lang. Processing, vol. 19, no. 6, pp. 1791-1801, August 2011.

[14] H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech," J. of Acoust. Soc. of Amer. (JASA), vol. 87, no. 4, pp. 1738-1752, 1990.

[15] H. Hermansky and N. Morgan, "RASTA Processing of speech ," IEEE Trans. on Speech and Audio Proce. vol. 2, no. 4, pp. 578-589, October 1994.

[16] Y. Steve et. al., The HTK book for HTK version 3.1.: Microsoft Corporation, July 2002.

[17] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in Proc. IEEE Workshop on Speech Synthesis, pp. 227-230, 2002.

[18] T. Yoshimura, K. Tokuda, T. Kobayashi, and T. Kitamura, "Siultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis system," in Proc. EUROSPEECH, Hungary, vol.5, pp. 2347-2350, 1999.

[19] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis system," in Proc. of International Conference on Accoustics, Speech and Signal Proceesing (ICASSP), Florida, USA, vol.3, pp. 1315-1318, June 2002.

[20] S. Imai, K. Sumita, and C. Furuichi, "Mel Log Spectrum approximation filter for speech synthesis," Trans. Institute of Electronic and Communication Engineers of Japan (IECEJ), vol. J66, no. A, pp. 10-18, 1983.

[21] H. A. Patil et. al., "A syllable-based framework for unit selection synthesis in 13 Indian languages," in Oriental International Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA) Conference, Gurgaon, India, 2013.

[22] [Online]. http ://hts.sp.nitech.ac.jp/ (Last Accessed on October 27, 2013).

[23] H. A. Patil, M. C. Madhavi, K. D. Malde, and B. B. Vachhani, "Phonetic transcription for fricatives and plosives for Gujarati and Marathi languages," in International conference on Asian Language Processing (IALP), Vietnam, pp. 177-180, 2012.

[24] Int. Telecom Union, "A method for subjective performance assessment of the quality of speech voice output devices," ITU-T Rec., P.85, 1994.

[25] DMOS. http :// www.itu.int/rec/ T-REC-P.800-199608-I/en (Last Accesed Octber 27th, 2013).

[26] C. Benoit, M. Griceb, and V. Hazanc, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," Speech Communication, vol. 18, no. 4, pp. 381-392, June 1996.

[27] R. F. Kubicheck, "Mel-cepstral distance measure for objective speech quality assessment," in Conference on Communications, Computers and Signal Processing, Victoria, BC, pp. 125-128,1993.

[28] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. on Acoust. Speech and Signal Process. , vol. 26, no. 1, pp. 43-49, February 1978.

[29] R. Boothalingam, et. al., "Development and evaluation of unit selection and HMM-based speech synthesis system for Tamil", National Conference on Communications (NCC), pp. 1-5, 2013.

[30] B. Ramani, et.al., "A common attribute based unified HTS framework for speech synthesis in Indian languages," 8th ISCA Speech Synthesis Workshop (SSW8), Barcelona, Spain, pp. 291-296, 2013.

[31] P. Daniel, W. Ellis, PLP, RASTA and MFCC in MATLAB, [Online]:http://www.ee.columbia.edu/~dpwe/resources/ matlab/rastamat/, (Last Accessed October 27th, 2013).