ANALYSIS OF NATURAL AND SYNTHETIC SPEECH USING FUJISAKI MODEL

Tanvina B. Patel and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), India
Gandhinagar-382007, India
{tanvina_bhupendrabhai_patel, hemant_patil}@daiict.ac.in

ABSTRACT

Text-to-speech (TTS) synthesis systems are being advanced to achieve *naturalness* and *intelligibility* in synthetic speech. Unit selection-based synthesis (USS) and Hidden Markov Model-based text-to-speech synthesis systems (HTS) are recent techniques in this area. USS-based synthetic speech is known to be natural (due to *concatenation* of natural speech sound units). On the other hand, HTS-based speech is *not* as natural in perception as USS-based synthetic speech. Due to speech synthesis technologies, voice biometrics systems may face threats due to impostor attacks. Thus, it is important to study the differences that exist between natural and synthetic speech. In this context, we investigate the effectiveness of parameters of *Fujisaki model* for capturing Fundamental frequency (F_0) contour variations in natural and synthetic speech. F_0 contour of speech contains linguistic and non-linguistic information. Experimental results on several utterances from Gujarati (a low resourced language) demonstrate the effectiveness of *phrase* and *accent* components to analyze the difference between these two speeches. Variability in phrase and accent components suggests that synthetic speech differs in terms of *prosodic* information in excitation source as compared to natural speech. These findings may assist to distinguish these two speeches and provide an aid to *alleviate* impostor attacks.

Index Terms-Fujisaki model, accent, phrase, synthetic speech.

1. INTRODUCTION

Text-to-Speech (TTS) synthesis techniques have remarkably developed in the past decade. One of these techniques includes development of unit selection-based synthesis (USS) using Festival framework [1] (which requires large recorded data and its transcription). USS synthesized speech is natural in perception and intelligible if the transcription is accurate enough. Another technique is statistical parametric speech synthesis based on Hidden Markov Model (HMM)-based TTS synthesis systems (HTS) [2], [3], [4]). Attempts to improve quality of speech synthesis techniques are being vulnerable for speaker recognition and verification systems due to the problem of impostor attack. Due to less training data required, HTS-based attacks are more common than the USS-based attacks. This occurs in scenarios where it is needed to authenticate if a particular speech is uttered by a genuine speaker. Other attacks may be due to mimicking (twins), replay, face under cover and voice transformation, etc. Here, we look into attacks due to state-of-the-art USS and HTS synthesis systems. In this paper, we propose to analyze natural *vs.* synthetic speech by exploiting prosodic information in excitation source during speech production.

Fundamental frequency (F_0) carries *linguistic* and *non-linguistic* information [5]. It carries information such as the speaker's identity, emotion, mood, etc. F_0 contour is known to be the result of movement of *intrinsic* muscles in the larynx [5]- [6]. The model representing these movements is well known as the *Fujisaki model* or the *command-response* model. It represents F_0 contour in terms of the *phrase* and *accent* parameters as a result of the *translation* and *rotation* motion of the cricoids muscles, respectively [7]. Earlier in [8]-[9] features like the pitch pattern and its variability were used as source features to detect synthetic speech. Other than source-based features, in [10] relative phase shift (RPS) was used to detect synthetic speech. Here, we use variations in F_0 contour (in terms of Fujisaki model parameters) of natural and synthesised speech to study the source-based discriminative features between the two speeches. In mimic speech, an imitator tries to vary his/her F_0 contour so that the shape of the F_0 contour matches with the target speakers F_0 contour [11]. However, speech synthesis technologies are not so human-like to perform such close matching to the F_0 contour of the natural speech. In other words, the natural speech is uttered with appropriate *breaks* and *accent* variations, which is not the case for synthetic speech. Therefore, we investigate the Fujisaki model parameters of natural and synthetic speech for discriminating these two speeches.

2. THE FUJISAKI MODEL

The Fujisaki model assumes the F_0 contour in log-domain as the *superposition* of two mutually independent contributions that occurs due to the independent movement of the thyroid cartilage and muscular reaction times [5]-[7]. Fujisaki model works for various languages (i.e., it gives good synthetic model contours) [6]. Here, we use Fujisaki model for Gujarati (a low resourced Indian language).

The authors would like to thank Department of Electronics and Information Technology (DeitY), New Delhi, India for TTS (Phase-II) and ASR (Phase-II) consortium projects and the authorities of DA-IICT.

2.1 The Model

Fujisaki model is a *superposition* of *phrase components* (y_p) and *accent components* (y_a). The two contributions are superimposed with a constant value F_b (i.e., minimum value of the speaker's F_0, which is known to be speaker-specific) to give a particular model generated F_0 contour as in Fig. 1.

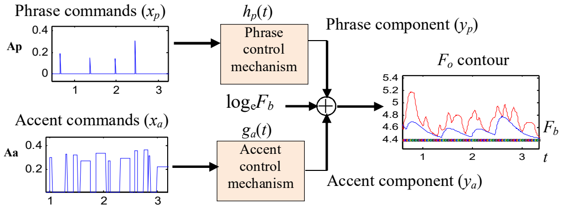

Fig. 1. Functional command response model for the process of generating F_0 contour. After [5], [7].

The phrase and accent components are *impulse* response and *step* response of critically damped 2^{nd} order linear system that is excited by Dirac impulses called *phrase commands* (x_p) and rectangular pulses called *accent commands* (x_a), respectively [5]. The output of phrase control mechanism (y_p) is characterized by the following impulse response,

$$h_p(t) = \alpha^2 t e^{-\alpha t} u(t), \tag{1}$$

where $\alpha \in [2,4]$ s^{-1} is its natural angular frequency. On the other hand, the output of the accent control mechanism (y_a) is characterized by following step response,

$$g_a(t) = [1-(1+\beta\,t)]e^{-\beta t}]u(t), \tag{2}$$

where $\beta \in [19,21]$ s^{-1} is its natural angular frequency. The total F_0 contour of the utterance can be expressed as,

$$y(t) = \ln[F_0(t)] - \ln[F_b] = y_p(t) + y_a(t)\,, \tag{3}$$

$$y(t) \;= \sum_{k=1}^{Np} A_{p,k} h_p(t - t_{p,k}) + \sum_{k=1}^{Na} A_{a,k}[g_a(t - t'_{a,k}) - g_a(t - t''_{a,k})], \tag{4}$$

where N_p and N_a are the number of phrase and accent events; $A_{p,k}$ and $t_{p,k}$ are the magnitude and timing of the k^{th} phrase command; $A_{a,k}$, $t'_{a,k}$ and $t''_{a,k}$ are the magnitude, onset time and the end time of k^{th} accent command. In Fig. 1, the nonlinear system for glottal airflow effects has been ignored [5].

2.2 Extraction of Model Parameters

2.2.1 F_0 Extraction

Extracting *prosodic* events from speech requires estimating F_0 contour of speech. Here, we use the zero frequency filtering (ZFF) method to estimate the F_0 contour [12]. The epoch locations, i.e., the glottal closure instants (GCIs) are obtained from the negative-to-positive zero-crossings of the zero frequency filtered signal. Thereafter, the F_0 contour is obtained from the GCI locations. Fujisaki model requires a continuous contour and it deals with *macroprosody* only. Hence, two tasks are performed before modeling the F_0 contour; (1) intermediate values for unvoiced speech regions and short pauses are *interpolated* in the F_0 contour, (2) microprosodic variations due to individual speech sounds units (such as plosive, fricatives, etc.) are smoothed out. Here, a linear fit is used during interpolation and then the F_0 contour has been smoothed prior to parameter estimation. Fig. 2 shows the F_0 contour extracted from ZFF algorithm where the unvoiced regions are interpolated with a linear fit.

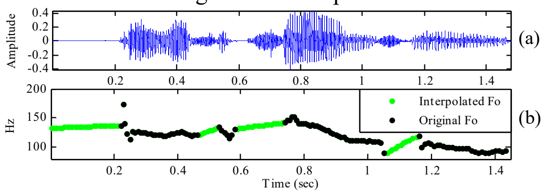

Fig. 2. (a) Speech from TIMIT database [13] (*16* kHz), (b) original F_0 contour (black) and linear interpolated F_0 contour (green).

2.2.2 Phrase and Accent Command Extraction

The processed F_0 contour is then used to estimate the phrase and accent events. The detection of phrase events or phrase boundaries is based upon the work reported in [14]. The F_0 contour is lowpass filtered and the negative-to-positive transition of the derivative of the filtered F_0 contour is taken as phrase boundaries. The *strength* of the phrase boundary was estimated by *slope* of the line at the negative-to-positive crossings. Accent commands parameter extraction is based on the work carried out in [15]. The procedure is to detect the largest maximum and the smallest minimum for each interval where the sign of the derivative of F_0 remains same. A pair of maximum and minimum corresponds to the *onset* and the *offset* of an accent command. An example of the Fujisaki parameters extracted from a speech utterance at *16* kHz is shown in Fig. 3. It is observed that model generated contour approximates smooth version of original F_0 contour.

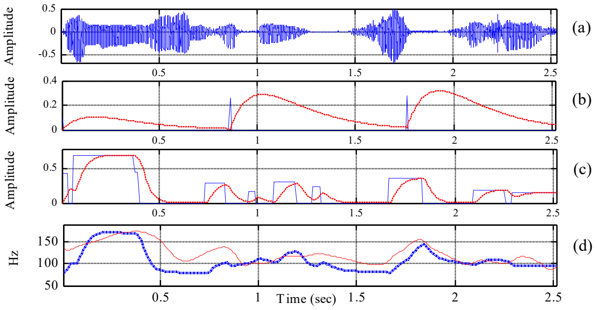

Fig. 3. (a) Speech signal, (b) phrase commands (blue) and phrase components (dashed), (c) accent commands (blue) and accent components (dashed) and (d) original F_0 contour and model generated F_0 contour (dashed).

3. GENERATION OF SYNTHETIC VOICES

The USS system is built using Festival framework [1], [16]. Since Indian languages are generally syllable-timed [17], we use syllable as the speech sound unit. Text optimization was carried to obtain *95 %* syllable coverage [16]. The speaker selection was done as in [18] and the recorded data was labeled using semiautomatic *DONLabel* labeling tool that uses group delay-based techniques for labeling at syllable-level [19]. The labeled data was manually aligned by trained professionals. The Gujarati USS systems (male and female) were built on *8* hours of speech data with an average mean opinion score (MOS) of around *3.3.* HMM-based synthesis framework gives a general setup for context modeling and is easily adapted to other languages [20]. For Gujarati the phonemic representation consists of *49* phonemes, broadly classified into SIL (i.e., silence), *36* consonants and *12* vowels [21]. HTS systems were built on *5* hours of speech data with average MOS of *3* (male) and *2.7* (female).

4. SPECTROGRAPHIC ANALYSIS

For experiments, *100* male and *100* female utterances were used each for natural, USS and HTS system. Same text material is used for both natural and synthetic speech. Fig. 4 shows the spectrogram for natural speech, USS and HTS-based synthesized speech for the same utterance. The spectrogram of USS-based synthesized speech is similar to the natural speech in terms of speaker characteristics. However, there are breaks in the spectrogram representing discontinuity in the formant contour (dotted oval showing abruptness due to concatenation). These breaks may also occur in natural speech however, the frequency of their occurrences in USS-based speech is relatively more due to concatenation of speech sound units. The spectrogram of HTS-based speech shows loss in intelligibility and the formant structure do not appear to be preserved in HTS-based speech (dotted squares).

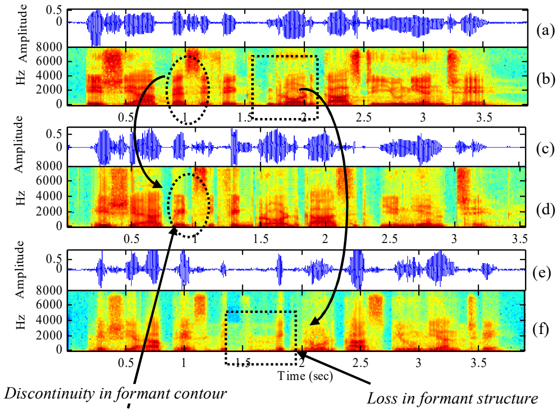

Fig. 4. (a) Speech Signal, (b) spectrogram of (a), (c) USS speech, (d) spectrogram of (c), (e) HTS speech and (f) spectrogram of (e).

We quantify the difference in various spectrograms by the *Itakura–Saito* (IS) distance measure. It measures *perceptual* difference between an original spectrum $P(\omega)$ and its approximation $\hat{P}(\omega)$. We consider synthetic speech to be an approximation to the natural speech. The utterances were time-aligned using *Dynamic Time Warping* (DTW) [22] and linear prediction coefficients (LPC) were extracted from the speech signal for every *20* ms speech frame with a frame shift of *10* ms for computation of IS distance, given by [23],

$$D_{IS}(P(\omega),\hat{P}(\omega)) = \frac{1}{N}\sum_{m=1}^{N}\left[\frac{P(\omega_m)}{\hat{P}(\omega_m)} - \log\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right) - 1\right], \qquad (5)$$

where *N* is the number of speech frames. Table 1 shows that the IS distance is relatively less between natural and USS speech as compared to HTS speech because IS distance measures spectral characteristics reliant on the *size* and *shape* of the vocal tract of the individual.

Table 1. Average IS distance between natural and synthetic speech over *100* male and *100* female utterances.

D_{IS}	USS		HTS	
	Male	**Female**	**Male**	**Female**
IS	11.675754	9.2565341	14.994685	18.448603

5. ANALYSIS BY FUJISAKI MODEL PARAMETERS

The model generated F_0 contour in log-domain is a superposition of phrase component (y_p), accent component (y_a) and F_b. The following sub-section shows variations of these parameters for natural and synthetic speeches.

5.1 Minimum Value of F_0 Contour (F_b)

In the representation of Fujisaki model, F_b is the baseline value of F_0 contour. It is a constant term $c_o(T_o/\sigma)^{1/2}$, where c_o is a constant *inversely* proportional to *size* of the membrane, T_0 indicates the *static tension* applied to the vocal fold and σ is the density per unit area of the membrane [7]. F_b is approximated as a constant as long as the speaker maintains *same* speaking style and emotional state [6]. F_b for female should be higher than male for natural, USS and HTS voices (as in Table 2). For USS, the speech sound units are concatenated using units from the same speaker. Therefore, mean value of their F_b is nearly same to natural. However, as the units are concatenated from several sessions of recording, there exists more variability in the F_b (more standard deviation (*sd*)). More the professional artists are consistent in recording, lesser would be the variation in F_b. For HTS, the mean F_b will be close to the natural speech if naturalness in HTS speech is persevered. However, other direct inferences for *sd* could not be drawn; hence, phrase and accents components are used for further analysis.

5.2 Phrase Commands and Phrase Components

The instant when the *cricoid* muscles undergo a translation motion, an *impulse* is generated corresponding to *phrase breaks* which occur naturally during speech production. Such prosodic breaks are not more prominent in synthetic speech. Fig. 5 shows the number of breaks in the natural and the synthetic speeches. In USS, silences are introduced in synthetic speech as per the text punctuations, whereas during natural speech production, prosody is automatically generated as per the nature of the utterance, context, etc. Therefore, USS synthesized speech is expected to have less

Table 2. The distribution (in terms of the mean and standard deviation (*sd*)) for *100* male and *100* female utterances (USS and HTS) of the minimum value of F_0 contour (F_b), the phrase components (y_p) and the accent components (y_a).

Parameters of model	Natural				USS				HTS			
	Male		Female		Male		Female		Male		Female	
	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd
F_b	64.79	6.0454	123.12	19.374	67.73	8.7885	131.97	22.313	65.51	10.0272	92.69	14.138
y_p	0.1158	0.1055	0.3894	0.2722	0.1481	0.1082	0.3146	0.2293	0.1792	0.1288	0.2456	0.1969
y_a	0.3396	0.2506	0.3753	0.2811	0.2797	0.1993	0.3403	0.2313	0.2909	0.1651	0.4266	0.3121

or equivalent number of *phrase* breaks as compared to the natural speech as seen in Fig. 5. For HTS male and female, the number of phrase breaks increases. Next, impulses due to phrase commands are passed through 2^{nd} order phrase control mechanism, to produce the phrase components. USS synthesized speech have similar means and *sd* for phrase components to that of natural utterances (as in Table 2). In HTS, for the number of phrase breaks was more. However, their strength was less due to the fact that these breaks were due to phrase pauses decided from text and not due actual change in translational motion of the cricoid muscles. Hence, mean value of phrase component is less.

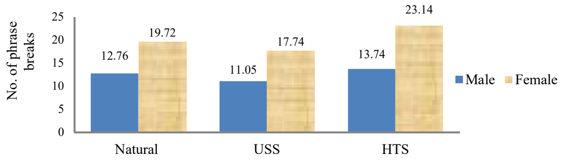

Fig. 5. Number of phrase breaks in natural, USS and HTS speech.

5.3 Accent Commands and Accent Components

The accent parameters of the Fujisaki model capture the variations in the speech which includes the stress that is applied to a particular word, syllable, etc. Especially for interrogative and exclamatory types of sentences, the accent parameters vary significantly. In case of USS speech, due to concatenation, the natural variation due to stress on syllable, word etc. may not always be present. Hence, the synthetic speech is monotonous in listening and this brings a possibility for the accent commands and components to vary less than natural speech. Thus, for USS synthesized speech, the mean and *sd* of the accent components was less than the natural speeches (as shown in Table 2). In case of HTS, any uniform pattern for the accent parameters was not found for either male or female.

5.4 Statistical Analysis of Results

The scatter plots for *100* USS and HTS synthesized voices formed by the mean of accent and phrase components are shown in Fig. 6. The clusters for USS and natural speech are different in size and shape than HTS and natural speech. In particular, clusters for USS synthesized and natural speech are found to be more overlapping and it is difficult to identify a boundary for these two classes. On the other hand, for HTS speech (especially female voice), the two classes are easily *separable*. Thus, the female voice in HTS *lacks* relatively the prosodic features as that of the natural voice.

To know the difference in distribution of the parameters for the natural and synthetic voices, we performed the Student's *t-test* and investigate if the two sets of synthetic voices are significantly different from natural. It is seen from Table 3 that the *null* hypothesis for all the synthetic voices is *rejected* with very less probability (<than *0.0001*) in most cases. Hence, the natural and synthetic systems (USS and HTS) have diverse means, which is effective while training statistical models like GMM, etc. This shows that the phrase and accent parameters could prove a good set of features to distinguish natural and synthetic.

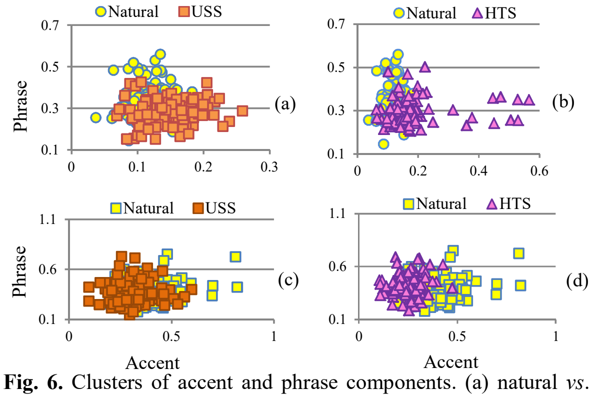

Fig. 6. Clusters of accent and phrase components. (a) natural *vs*. USS (male), (b) natural *vs*. HTS (male), (c) natural *vs*. USS (female) and (d) natural *vs*. HTS (female).

Table 3. Probability of rejecting null hypothesis for phrase and accent parameters in USS and HTS.

System	USS		HTS	
	Phrase	Accent	Phrase	Accent
M	<0.0001	<0.0001	<0.0001	<0.0001
F	<0.0001	0.018	<0.0001	0.0002

6. SUMMARY AND CONCLUSIONS

This paper presented an analysis of natural *vs*. synthetic speech using phrase and accent parameters of the Fujisaki model. For both USS and HTS, it was observed that phrase command could serve as an important feature to distinguish between natural and synthetic speech. For HTS speech, phrase components could be effective. Results of *t-test* show that the null hypothesis is rejected in all cases, which is effective while training statistical models on large dataset. Thus, our future research will be directed towards exploring these excitation source features for classification problem and apply these features to alleviate voice biometric attack.

7. REFERENCES

[1] A. Black, P. Taylor and R. Caley, "The Festival speech synthesis system," 1988. [Available Online]: http://festvox.org/festival/ {Last accesssed: *24th* Aug. 2015}.

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

[3] H. Zen, K. Tokuda and A. W. Black, "Statistical parametric speech synthesis," *Speech Comm.,* vol. 51, no. 11, pp. 1039-1064, Nov. 2009.

[4] J. Yamagishi, T. Nose, H. Zen, Z. -H. Ling, T. Toda, K. Tokuda, S. King and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio and Lang. Process.,* vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[5] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretations," Dept. for Speech, Music and Hearing, KTH Stockholm, Quarterly Progress and Status Report, vol. 2, no. 1, pp. 1-20, 1981.

[6] H. Fujisaki, S. Ohno and W. Gu, "Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command–response model for generation of their F_0 contours," in *Proc. of Int. Symposium on Tonal Aspects of Lang.—with Emphasis on Tone Lang.*, pp. 61-64, Beijing, China, 2004.

[7] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in *Proc. of Speech Prosody*, Nara, Japan, pp. 1-10, March 2004.

[8] A. Ogihara, H. Unno and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Trans. on Fundamentals of Elect. Comm. and Computer Sciences,* vol. 88-A, pp. 280-286, 2005.

[9] P. L. De Leon, B. Steward, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Portland, Oregon, USA, pp. 370-373, 2012.

[10] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Process., (ICASSP),* Prague, Czech Republic, pp. 4844-4847, 2011.

[11] D. Gomathi, S. A. Thati, K. V. Sridaran and B. Yegnanarayana, "Analysis of mimicry speech," in *Proc. Int. Speech Comm. Assoc. (INTERSPEECH)*, Portland, Oregon, pp. 695-698, 2012.

[12] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. on Speech and Audio Process.,* vol. 16, no. 8, pp. 1602-1613, Nov. 2008.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. Dahlgren and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," Philadelphia: Linguistic Data Consortium, 1993.

[14] A. Sakurai and H. Hirose, "Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours," in *Proc. Int. Conf. on Spoken Lang. Process., (ICSLP)*, Philadelphia, PA, vol. 2, pp. 817-820, 1996.

[15] S. Narusawa, N. Minematsu, K. Hirose and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in *Proc. IEEE Int. Conf. Acoust., Speech and Sig. Process., (ICASSP)*, Orlando, Florida, USA, pp. 509-512, 2002.

[16] H. A. Patil et. al., "A syllable-based framework for unit selection synthesis in *13* Indian languages," in *Proc. Oriental Int. Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA) Conference*, Gurgaon, India, pp. 1-7, 25th -27th Nov. 2013.

[17] K. S. Prahallad and A. W. Black, "Unit size in unit selection speech synthesis," in *Proc. Eur. Conf. Speech Comm. Technol. (EUROSPEECH)*, Geneva, pp. 1317-1320, 2003.

[18] H. Patil, T. Patel, S. Talesara, N. Shah, H. Sailor, B. Vachhani, J. Akhani, B. Kankariya, Y. Gaur and V. Prajapati, "Algorithm for speech segmentation at syllable-level for text-to-speech synthesis system in Gujarati," in *Proc. Oriental Int. Committee for the Co-Ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA) Conference*, Gurgaon, India, pp. 1-7, 2013.

[19] P. G. Deivapalan, M. Jha, R. Guttikonda, and H. A. Murthy, "DONLabel: Automatic labeling tool for Indian languages," in *Proc. National Conf. Comm. (NCC)*, IIT Bombay, India, pp. 263-266, 2008.

[20] "Nagoya Institute of Technology," [Available Online].: http://hts.sp.nitech.ac.jp/ {Last accessed: *24th* August 2014}.

[21] 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 *Proc. Int. Conf. on Asian Lang. Process., (IALP)*, Hanoi, Vietnam, pp. 177-180, 2012.

[22] R. F. Kubicheck, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. on Comm., Comp. and Sig. Process.*, Victoria, BC, pp. 125-128, 1993.

[23] S. R. Quackenbush, T. P. Barnwell and M. A. Clements, Objective measures of speech quality, Prentice-Hall, Eaglewood Cliffs, 1988.