ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Feature Extraction and Analysis of Speech Quality for Tamil Text-To-Speech Synthesis System using Fast Fourier Transform

[1]K.C.Rajeswari and [2]Dr.P.Uma Maheswari

[1]*Anna University, CSE Department, Sona College of Technology, Salem, Tamil Nadu, India.*
[2]*Anna University, CSE Department, -Madras Institute of Technology Campus, Chennai, Tamil Nadu, India.*

ARTICLE INFO

Article history:
Received 1 November 2015
Accepted 28 November 2015
Available online 18 January 2016

Keywords:
F0 extraction, intonation, mel-frequency cepstral coefficients, peak analysis, speech synthesis, Tamil TTS

ABSTRACT

Speech is an art and unique in nature when pronounced by the human beings. Whereas in technical terms, speech is a signal that is generated while vocal tract is energized by the impulses of the air that comes from the lungs through the vocal cords. The quality of the speech influences the precision and pleasure of the hearer. Speech processing techniques are intended to improve the speech quality while intelligibility and naturalness is not ample. These days, as voice-enabled applications are trendy, researchers focus primarily on speech processing techniques to provide a high quality speech enabled devices and applications. It is also important to develop such applications in regional languages for privileged usability. The human community feels immense pleasure when auditory sense is completely satisfied. The Information and Communication Technology enabled devices and products are widely used by everyone but lack of high quality speech facilities in the devices still exist. In this work, one of the speeches processing tasks Text-to-Speech synthesis is dealt. Essential attributes of both text and speech components are extracted and analyzed using Fast Fourier Transform to improve the quality of speech. The experimental results have shown significant improvement in speech quality.

To Cite This Article: K.C.Rajeswari and Dr.P.Uma Maheswari., Feature Extraction and Analysis of Speech Quality for Tamil Text-To-Speech Synthesis System using Fast Fourier Transform. ***Aust. J. Basic & Appl. Sci.,*** *9(35): 349-356, 2015*

INTRODUCTION

In India, there are 29 states, in which 74 million people are living and 16 million people are illiterates. Moreover, people in India speak 1652 dialects but 17 languages are declared as official languages by the Indian Constitution. To facilitate the people by all means, Government of India spend enormous money for Language Research and Information and Communication Technological development. In spite of all these, still there is a huge demand for quality enhanced speech products**.** Text-to-Speech synthesis is one of the speech processing task which takes arbitrary text as input and produce speech as output. Text-to-Speech Synthesis systems are being developed in many languages. Hema A Murthy and *et al* (2010) mentioned that the Government of India has taken initiatives to develop TTS for all Indian languages by forming a consortium in which IIT-Madras, IIT-Kharagpur, IIIT-Hyderabad C-DAC-Mumbai and C-DAC-Trivandrum are the members. IIIT, Hyderabad has developed a TTS engine for Telugu language and contributed screen readers for visually challenged people. Jayasankar *et al*., (2014) developed a channel vocoder and proposed cochlear implant for analyzing and synthesizing the Tamil words. TTS are available for Hindi and Bengali languages. IISC-Bangalore also progress the research for developing Speech synthesizer in Kannada and Tamil language.Each of these synthesizers' has unique features but still differ in synthesis approach.

The various approaches include formant based synthesis ,concatenation based synthesis and articulatory synthesis. Formant synthesis makes use of the formant transitions present across vowels and consonants. Articulatory synthesis makes use of different aspects and configuration of speech organs when the speech is produced and simulates the same when speech is synthesized. Concatenative synthesis makes use of recorded speech samples, segment them into phones, di phones or syllables and process the samples to finally concatenate the segments as it produces synthetic speech. Youcef Tabet *et al* (2011) in their research have proven that concatenative synthesis is comparatively good to yield natural sounding and intelligible speech. Jayasankar *et al* (2014)proposed FPGA based concatenative Text to Speech synthesis system in which text analysis and syllable preparation is possible that will improve naturalness in speech. This is unique because speech

Corresponding Author: Ms. K.C.Rajeswari, Anna University, CSE Department, Sona College of Technology, Salem- 636 005, TamilNadu, India.
Tel: +91 9600516927; E-mail: rajeswarikc@sonatech.ac.in

quality improvement techniques used so far is only using software. Khalifa *et al* (2011) developed a rule based Arabic speech synthesis system using hybrid synthesis technique in which phonemes are the units produced reasonably good quality of speech. In all these approaches, producing high quality speech is of special interest and it is the work of concern. When the quality of speech is of special issue, researchers are trying to come up with different feature extraction techniques, perfect training of data, nature of input, type of neural network used for training and testing the data. The prosodic parameters namely duration, intonation and intensity are responsible for producing high quality speech. Especially, intonation is the component one has to consider because it is the most expressive portion of the speech. Rajeswari *et al* (2012), Jayavardhana Rama G L *et al* (2002) and Ramu Reddy *et al* (2013) stated that the intonation must be used for feature extraction as Fundamental frequency (F0) if analyzed and used, there is a prospect to bring naturalness in the speech produced. In this paper, feature extraction of text and audio are presented in section II, the application of Fast Fourier transform is explained section III, section IV shows the experimental setup preferred for the proposed work along with the results, and finally section V concludes the work.

II .Feature Extraction:

Since the extraction of features plays a vital role in the training and testing phase of the system, the existing TTS systems incorporated many techniques for feature extraction .In general, text features are extracted from the words, characters, spaces and special characters, etc. Ramu Reddy *et al* (2012) used the positional, contextual, phonological and articulatory features to predict the F0 contours and the TTS was developed for Bengali language. The production constraints are represented using articulatory features but it is very tedious process. The study of production of human speech sounds at different articulatory positions is very complex. The clear understanding of movement of speech organs is also a prerequisite to extract the intonation features. Sreenivasa Rao (2009), Antonis Botinis (2001) and Gopala Krishnan *et al* (2011) has concluded from their research that intonation is the parameter that describes the dynamics of fundamental frequency F0 and it directly correlates to the pitch. Such an important factor cannot be ignored; instead there arises a need for proposing a novel technique to extract the features. The proposed system makes use of intonation functions instead of articulatory features where intonation functions ensure the promising results in feature extraction with minimal effort and simple approach. Intonation is the component that better conveys the pragmatics and emotions in the speech produced. In the proposed system, input data is transliterated form of Tamil text and can be stored in a file and measured length of the input text is stored separately. The type of the sentence is identified as whether declarative, interrogative, exclamatory using the special characters and spaces present in the sentence. In the input sentence, syllables are taken and its corresponding ASCII code is stored. Once the type of the sentence is identified, the labels are assigned its corresponding values. The amplitude of the speech signal corresponds to pitch. Here, nine feature vectors are used because the proposed system not only extracts the amplitude but also gathers positive and negative peak variations of the pitch. During text feature extraction, index values are calculated and used when the user, enter the text of their own instead of stored data. For training purpose, 80% of the features are used. The features are stored in 9 feature vectors namely max (R_loc), max (Q_loc), max (S_loc), length (R_loc>0), length (Q_loc>0), length (S_loc>0), sum (R_loc>0), sum (Q_loc>0) and sum(S_loc>0). These features are used for training purpose initially and help in normalizing the signal. The convolution process is carried out at every stage as the transfer function is applied. Finally, a centroid threshold is obtained which is very important to extract the immediate left and right positive and negative peak variations of the pitch. The step by step procedure illustrated below shows the process of obtaining updated threshold limitation. Threshold is not a random value; to validate the threshold, the updating is done.

$$X_1 = \frac{S_1}{\max(S_1 - mean(S_1))} \text{----------} \quad (1)$$

$$X_2 = \frac{\left((\sum_{m=0}^{N}(x_1(m)*h_1(m)))*(T_s)\right)}{\left(max\left((\sum_{m=0}^{N}(x_1(m)*h_1(m)))*(T_s)\right)\right)} \text{------} \quad (2)$$

$$X_3 = \frac{\left((\sum_{m=0}^{N}(x_1(m)*h_1(m)))*(T_s)\right)}{\left(max\left((\sum_{m=0}^{N}(x_1(m)*h_1(m)))*(T_s)\right)\right)} \quad (3)$$

$$X_4 = \frac{\left((\sum_{m=0}^{N}(x_1(m)*h_1(m)))*(T_s)\right)}{\left(max\left((\sum_{m=0}^{N}(x_1(m)*h_1(m)))*(T_s)\right)\right)} \quad (4)$$

$$X_5 = \frac{(X_4)^2}{max(X_4)} \text{----------} \quad (5)$$

$$X_6 = \frac{\left((\sum_{m=0}^{N}(x_1(m)*h_1(m)))*(T_s)\right)}{\left(max\left((\sum_{m=0}^{N}(x_1(m)*h_1(m)))*(T_s)\right)\right)} \text{-----} \quad (6)$$

In this feature extraction stage, R_Loc represents feature at which the wave is in high peak Positive and Q_Loc represents features at small signal difference at negative edge of the audio signal and S_Loc represents feature values at maximum signal difference at negative point of input audio signal. For each stage of convolution process, Low pass filter and High pass Filter are used for calculating difference in peak extraction using transfer function as represented by X1, X2, … X6 as in equs. (1) through (6) and this is updated by the threshold value extracted from input audio signal.In the above equations, x1 is the input speech signal, Ts is the sampling time and N is the number of samples of the signal. $h_1(m)$, $h_2(m)$and $h(m)$are the transfer function of the low pass filter, high pass filter and

convolution respectively. The feature vectors are expressed as given in eqs(7),(8) and (9):

$R_{loc} = indx_{max}\left(X_{1(left \rightarrow Right)}\right) - 1 + Left$ --- (7)

$Q_{loc} = indx_{min}\left(X_{1(left \rightarrow R_{loc})}\right) - 1 + Lef$ ---- (8)

$S_{loc} = indx_{min}\left(X_{1(left \rightarrow Right)}\right) - 1 + Left$--- (9)

Where,

Left $=\{Pos_{Reg} == 1\}$

Right $=\{Pos_{Reg} == -1\}$

$Pos_{Reg} = \{X_6 > (Thres * Max_h)\}$

$Thres = Mean(X_6)$

$Max_h = Max(X_6)$

Here Left and Right specifies the left position and right position of sampled input signal. Figure 1 shows the input speech signal as a .wav file and the speech signal after filtering the noise.

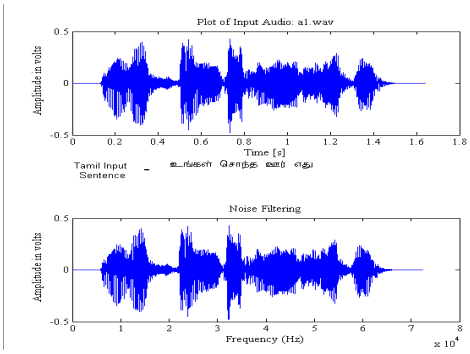


Fig. 1: Plot of Input speech signal and filtered signal:

III .Applying Fast Fourier Transform:

The input speech signal is sampled for discretization. In the next step, Fast Fourier transform shown in eq.(10) is applied over the output of the sampling process. After transformation is done, corresponding Mel-Frequency Cepstral Co-efficients (MFCC) are observed.

$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{i2\pi Kn}{N}}$ ---------- (10)

Where K=0,..,N-1

N = Sample size of audio signal x_n- Input Signal

The transformed discrete time signal is a conversion from time domain into its frequency domain. Rekha Hibare *et al* (2014) and Ahmed *et al* (1996) mentioned FFT is very fast and efficient to compute. The unique features of the speech signal can be extracted and analyzed with the help of MFCC. Rekha Hibare *et al* (2014) and Anjali Bala *et al* (2010) mentioned that MFCC clearly represent the short term power spectrum of the speech unit and also give log power spectrum of non- linear Mel scale. Suma swamy *et al* (2013) described that MFCC conveys the clear variation of the human auditory sense with the available critical bandwidth frequencies; linearly spaced filters at low frequencies and logarithmically at high frequencies. MFCC has its significance not only in extracting features for synthesis but aso for recognition is proven by Selva Vaidhyananthan *et al* (2014). The input speech signal is segmented into frames. Frame size chosen is 20 from the sample size of the input signal. Yao Qian *et al* (2011) suggested that in Fourier transformation, Discrete Cosine transform implemented is described as,

$C_n = \sum_{K=1}^{N}\left(\log X_k \cos\left(n\left(k - \left(\frac{1}{2}\right)\right) * \left(\frac{\pi}{k}\right)\right)\right)$ -- (11)

where n = 1,2,…N and N=sample size of speech signal input.

During Windowing, all the closest frequencies are gathered for feature extraction. It is observed that the window size if less than 25, it is noisier and if it is greater than 25, then losses is there. The window size is fixed as 25 and windowing is carried out which is described as,

$P(k) = \left(\frac{1}{N}\right)|X_I(k)|^2$ - (12)

In the MFCC based audio feature extraction, Mel Cepstrumis extracted from transformation output. In the next step, DFT is applied for Mel frequency warping and the signal in Hertz is converted to Mel warping using the following eq.(13).

$F(Mel) = (1127 * \log(1 + hz/700))$----- (13)

Inverse DFT is applied to extract the Mel Cepstrum output as feature of audio signal from MFCC. The resultant of this conversion is calledMel Frequency Cepstral Coefficients

Figure 2 shows the FFT transformed speech signal, Short term power spectrum and finally Mel-frequency cepstrum. It is necessary to obtain the deviations in the signal so that, it is helpful to extract the test features. The following figures shows the signal after DC drift cancellation and normalized output, signal after derivative, integrated output and signal with peak points extracted.

Interpolated αpeak location is given as,

$P = \left|\left(\frac{1}{2}\right) * \left(\frac{(\alpha-\gamma)}{\alpha-2\beta+\gamma}\right)\right|_{\left[\frac{-1}{2}, \frac{1}{2}\right]}$ ---------- (14)

The peak magnitude estimate is given as,

$T_1(p) = \beta - \frac{1}{4}(\alpha - \gamma)p$ ---------- (15)

Where,

α – Starting edge of parabola of the signal.

Australian Journal of Basic and Applied Sciences, 9(35) November 2015, Pages: 349-356

β- Peak amplitude edge of Signal.
γ - Finishing edge of parabola of the signal.

In this Peak Analysis, peaks are calculated from amplitude and frequency of input signal parameters α, β, and γ which can be calculated as from 'P' variable and check the condition of peak from peak magnitude as a threshold of peak estimation which indicates if 'P' is greater thanT1, then it is noted as peak range in that sampled size of signal.

Interpolated αpeak location is given as,

$$P= \left|\left(\frac{1}{2}\right) * \left(\frac{(\alpha-\gamma)}{\alpha-2\beta+\gamma}\right)\right|_{\left[\frac{-1}{2}\ ,\ \frac{1}{2}\right]} \text{----------} \quad (14)$$

The peak magnitude estimate is given as,

$$T_1(p) = \beta - \frac{1}{4}(\alpha-\gamma)p \text{----------} \quad (15)$$

Where,

α – Starting edge of parabola of the signal.
β- Peak amplitude edge of Signal.
γ - Finishing edge of parabola of the signal.

In this Peak Analysis, peaks are calculated from amplitude and frequency of input signal parameters α, β, and γ which can be calculated as from 'P' variable and check the condition of peak from peak magnitude as a threshold of peak estimation which indicates if 'P' is greater thanT1, then it is noted as peak range in that sampled size of signal.

.

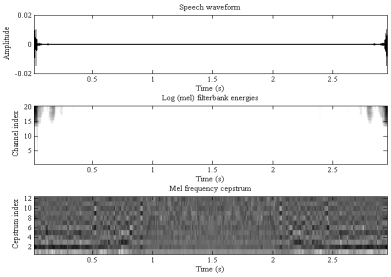
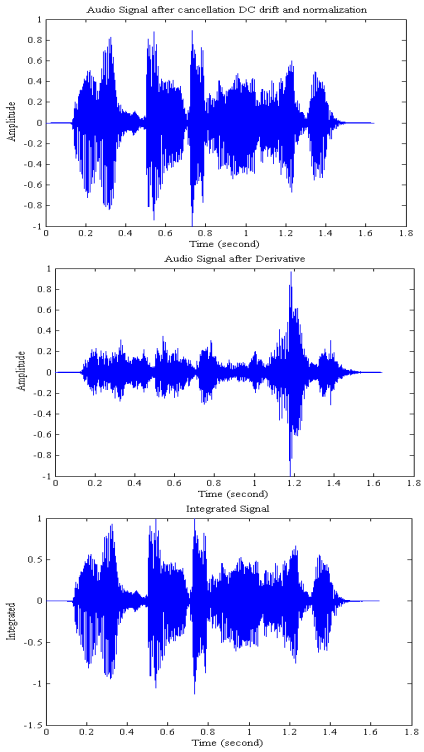


Fig. 2: Mel Frequency Cepstral Coefficients.



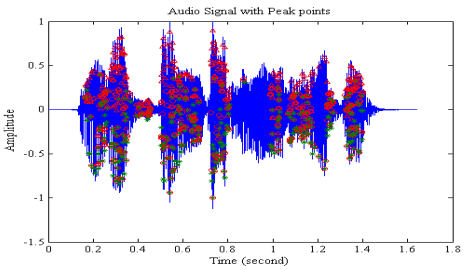


Fig 3: Signal after Dc drift cancelation and normalization, after derivative, integrated signal and signal with peak points.

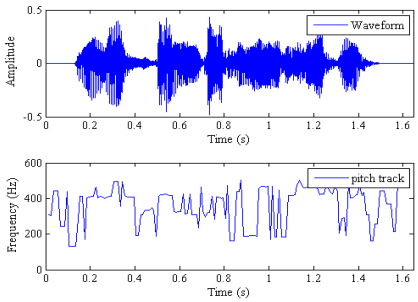


Fig. 4: Pitch track of the signal.

Here the pitch of signal is estimated using time domain based detection method with reference to Narendra N *et al* (2015). There are several methods available to estimate the pitch of the signal such as Zero Crossing, Autocorrelation, Maximum Likelihood, Adaptive filter using FFT and Super Resolution pitch detection. Among these, Maximum Likelihood based Pitch extraction is used which can be represent as,

$$\text{Pitch}(t,\tau)=\begin{Bmatrix}\frac{1}{N+1}\sum_{n=0}^{N}P(t+n\tau) & 0\le t\le b\\ \frac{1}{N}\sum_{n=0}^{N}P(t+n\tau) & b\le t\le\tau\end{Bmatrix} \text{- (16)}$$

where,

τ- Frame size of audio signal

t – Sampling time

N – Total size of audio signal.

The updation using the objective function is,

$$P(t+n\tau)=10*size(X_i)+\sum_{i=1}^{N}X_I^2-(10*\cos(2*pi*X_I))\text{ - (17)}$$

In this pitch extraction, first the objective function is implemented to perform weight calculation from the input speech signal based on cosine angle difference of the signal amplitude. Then from that objective function result, pitch angle variation for each pre-allocated time samples are extracted which is calculated from the length of input signal (Xi). Then the difference in limitation of time sequence with the Pitch(τ,T) calculation is verified and extracted the pitch of frequency difference in signal.

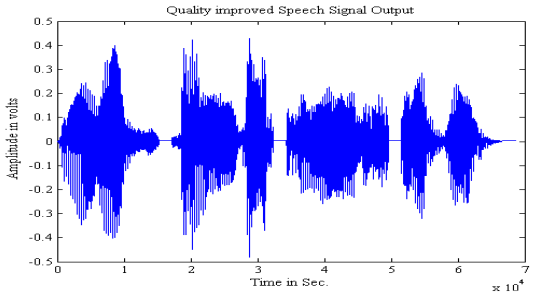


Fig. 5: High quality output speech signal.

In order to classify the extracted audio feature Neural Network is used as mentioned by Hung-Yan HU *et al* (2010) and Jainath Yadav *et al* (2015). Levenberg-Marquardt method is implemented and Gradient vector of proposed neural network is described as,

$$g_i=\sum_{p=1}^{P}\sum_{m=1}^{M}\left(\frac{\partial X_{p,m}}{\partial t_i}e_{p,m}\right)\text{----------}\quad\text{-(18)}$$

Where, e – feature vector of input signal, and

Xp,m- represents Feature matrix of Dataset.

In this classification stage, both text feature and audio features are supplied as input and retrieved the category of input text and speech signal to specify the width of speech signal to clarify the place where absence of signal can be experienced and where to raise the F0 to improve the quality of the signal is done.

From the classified result of neural network, it is obtained the class result of separation of speech signal to include level of amplitude with the help of pitch extraction. The class label represents marking of rising and falling range of speech input at particular sampling time. This is represented as,

$$Y(t+1) = Y(t) + Y(g_i) + Pitch(t,\tau)\text{------} \quad (19)$$

IV .Experimental setup:

Objective Evaluation:

The signal to Noise ratio and normalized correlation coefficients are the two measures used to analyze the performance of the proposed system. The SNR value of the given speech signal is 33.716. This value is quite high when compared to other systems. Similarly, the average normalized correlation coefficients are calculated and observed that average NCC is 10% more when compared to K-nearest neighbor method, Expectation-Maximization method, variation Bayes method and STOI method adopted by Jalal Taghia *et al* (2014) and Ke Hu *et al* (2013). The proposed system is subjected to test under various circumstances. It was analyzed how the SNR varies for different types of sentences namely interrogative, exclamatory, neutral, commanding, negotiation, interrogative special (type of interrogation). Its corresponding SNR values and functions of intonation for each type of sentences are depicted in the table3. The proposed system extract start, middle and end F0 as features, then analyzed the prediction of syllables within the deviation and observed that, if prediction level increases, then average prediction error will be minimized compared to Ramu Reddy *et al* (2013). It is shown in the table1 and 2 start, middle and end F0 values increases significantly when PCPF is used as the feature extraction. The intonation model CART in Festival features, Linear Regression using PCPA features, CART using PCPA, FFNN using PCPA are all less than the proposed system's PCPF features with 5 % and 15% prediction of syllables within the deviation.

Table 1: 5% and 15% Prediction of syllables within the deviation.

5% Predicted syllables within the deviation				15% Predicted syllables within the deviation		
Model	Start	Middle	End	Start	Middle	End
CART (Festival)	14.96	17.67	10.93	43.29	47.1	41.16
LR (PCPA)	21.32	29.17	19.97	61.5	78.93	59.99
CART (PCPA)	27.15	35.74	23.99	70.5	77.73	68.41
FFNN (PCPA)	31.26	40.58	23.18	73.16	80.81	69.89
Proposed(PCPF)	43.34	76.67	30	80	93.33	80

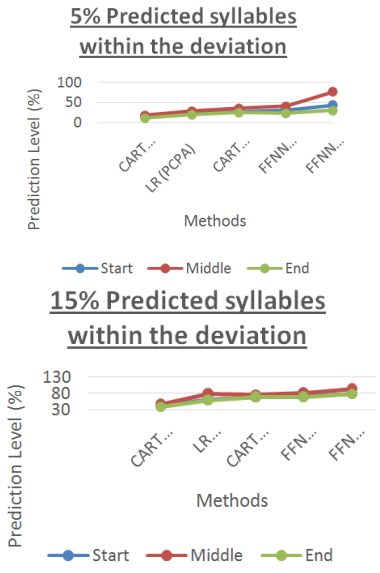


Fig. 6: Plot shows the variations between5% and 15%prediction levels of syllables within deviation when CART (Festival), LR, CART, FFNN (PCPA)and FFNN(PCPF) are used.

Subjective Evaluation:

The subjective evaluation is conducted in a sound proof environment. The evaluation is carried out for different speech signals by conducting two listening tests Test 2, Test 3 and Mean opinion score results are obtained from two categories of people. Test 2 is carried out with the subjects who are native Tamil speakers and Test 3 with the subjects who are

non-native Tami speakers. The table 2 shows the MOS values obtained in Test 2 and 3 for the different type of sentences. Brach Manski *et al* (2006) mentioned that MOS values are important for assessing the quality of speech. The proposed technique show considerable improvement in the quality of speech in terms of naturalness.

Table 2: Mean opinion scores of the quality improved speech.

Type of sentence	Intonation function	Test 1 SNR	Test 2 MOS	Test 3 MOS
Interrogative sentence	Grammatical function	23.176	3.72	3.91
Exclamatory sentence	Attitudinal function	11.9518	3.25	3.54
Neutral sentence	Psychological function	27.035	4.75	4.52
Commanding sentence	Indexical function	24.78	4.50	4.25
Negotiation	Discourse function	23.2688	4.21	3.56
Interrogative special	Focus function	23.215	4.16	3.89

Conclusion:

The proposed method uses intonation function as one of the text feature that improves the quality of speech. The input speech is subjected to Fast Fourier Transform, MFCC are generated, and variations at the peak are observed in section III Figure No. 2. The pitch F0 (intonation) is extracted as an important feature which improves the naturalness of the speech synthesized. The main advantage of the proposed model is the eradication of production constraints in feature extraction. The significant improvement in the quality of synthesized speech is shown in the Table 1.The proposed system will help the society in a wider range of applications such as talking tourist aid, voice enabled GPS in native language for the automobiles, screen readers for visually challenged people, people with speaking inability, voice enabled toys as a resource for educating kids at pre-schools, etc. The system can be further analyzed using Wavelet transform and Fractional Fourier transform for achieving intelligibility in speech.

REFERENCES

Aguero, P.D., K. Wimmer and A. Bonatonte, 2004. “Automatic analysis and synthesis of Fujisaki’s intonation model for TTS”, Speech Prosody.

Ahmed, I., Syed, 1996. “On the Relationship between the Fourier and Fractional Fourier Transforms”, IEEE Signal Processing, 3-1.

Anjali Bala, Abhijet Kumar and Nidhika Birla, 2010. “Voice Command Recognition System based on MFCC and DTW”, in the International Journal on Engineering Science and Technology, 2(12): 7335-7342.

AntonisBotinis, Bjorn Granstrom and Bernd Mobius, 2001. “Development and paradigms in Intonation Research”, Speech Communication, 33: 263-296.

BrachManski, S., 2006. “Experimental Comparison between Speech Transmission Index and Mean Opinion Scores in Rooms”, Archives of Acoustics, 31 (4): 171-176.

Gopala Krishnan A. and Luis C Oliveira and Alan W Black, 2011. “A Statistical Phrase/Accent Model for Intonation Modeing”, InterSpeech, 28-31.

Hema, A., Murthy, K.S. Rao and Kishore S. Prahallad, 2010. “Building Unit Selection Speech Synthesis in Indian Languages: An initiative by an Indian consortium”, in the proceedings of COCOSDA, Kathmandu, Nepal.

Hung-Yan, H.U., Ming-Yen LAI, and Sung-Feng TAI, 2010. “Combining HMM Spectral Models and ANN Prosody Models for Speech Synthesis of Syllable Prominent Languages”, in IEEE Transactions on audio, Speech and Language Processing, 451-457.

JainathYadav and SreenivasaRao, 2015. “Prosodic Mapping Using Neural Networks for Emotion Conversion in Hindi Language”, Circuits Syst Signal Process, DOI 10.1007/s00034-015-0051-3.

Jalal Taghia, and Rainer Martin, 2014. “Objective Intelligibility Measures Based on Mutual Information for Speech subjected to Speech Enhancement Processing”, in IEEE Transactions on audio, Speech and Language Processing, 22(1): 6-16.

Jayasankar T. and Arputha vijayasevi, 2014. “FPGA-based Implementation of text analyser and syllable preparation for concatenative Speech synthesis for Tamil Language”, in Australian Journal of Basic and Applied Sciences, 8(10): 102-109.

Jayasankar, T., Arputha Vijayaselvi, K. Rajasekaran and J. Vijayalakshmi, 2014. “Tamil Words Speech synthesis in cochlear implant using acoustic model”, in Australian Journal of Basic and Applied Sciences, 8(10): 349-356.

Jayavardhana Rama, G.L., A.G. Ramakrishnan, R. Muralishankar and R. Prathibha, 2002 “A Complete Text-To- Speech Synthesis System in Tamil”, in 0-7803-7395-2/02, IEEE proceedings of ICASSP.

Ke Hu and De Liang Wang, 2013. “An Unsupervised approach to Cochannel Speech Separation”, in IEEE Transactions on audio, Speech and Language Processing, 21(1): 122-131.

Khalifa, Othman Omran, Obaid, M.Z. Naji, Ahmed Wathik, and Daoud, Jamal Ibrahim, 2011. “A rule based Arabic Text-to-Speech system using hybrid synthesis technique” in Australian Journal of Basic and Applied Sciences, 5(6): 342-354.

Narendra, N. and K. Sreenivasa Rao, 2015. “Robust Voicing Detection and F0 Estimation for

HMM-based Speech Synthesis", Circuits Syst Signal Process, DOI 10.1007/s00034-015-9977-8.

Ramu Reddy, V. and K.S. Roa, 2013. "Two-stage intonation modeling using feed forward neural networks for Syllable based Text-to-Speech Synthesis," in Computer Speech and Language, 1105-1126.

Rajeswari, K.C. and P. UmaMaheswari, 2012. "Prosody Modeling Techniques for Tamil Text to Speech Synthesis Systems-A Survey," J. Computer Applications, vol. A247, 529-551.

Rajeswari, K.C. and P. UmaMaheswari, 2014. "A Novel Intonation Model to improve the quality of Tamil Text-to-Speech Synthesis Systems", in the proceedings of sixth International Conference on Advanced Computing, ICoAC, 335-340.

RekhaHibare and AnupVibhute, 2014. "Feature Extraction Techniques in Speech Processing: A Survey", International Journal of Computer Applications, 107-5.

Selva Vaidhyananthan, S., R. Shantha Selva kumari and T. Senthur Selvi, 2014. "Text independent voice based attendance system under multi speaker and noisy environment", in Australian Journal of Basic and Applied Sciences, 8(13): 121-131.

Sreenivasa Rao, K. and B. Yegnanarayana, 2009. "Intonation Modeling for Indian Languages", in Computer Speech and Language 23: 240-56.

Suma Swamy and K.V. Ramakrishnan, 2013. "An efficient Speech Recognition System", in Computer Science and Engineering: An International Journal, 3(4): 21-27.

Yao Qian, Zhizheng Wu, BoyangGao, and Frank K. Soong, 2011. "Improved Prosody Generation by Maximizing Joint Probability of State and longer Units", in IEEE Transactions on Audio, Speech and Language Processing, 19(6): 1702-1710.

YoucefTabet and Mohamed Boughazi, 2011. "Speech ynthesisTechniques.A survey", in the proceedings of 7th International workshop Systems, Signal Processing and their Applications(WOSSPA), 67-70.