VITAL METHOD FOR PITCH EXTRACTION IN HMM BASED SPEECH SYNTHESIS

[1]D. Kamala, [2]Dr. I. Santi Prabha
[1]P.G Scholar, Dept of ECE, University College of Engineering Kakinada, JNTUK, Kakinada, AP, India.
[2]Professor, Dept of ECE, University College of Engineering Kakinada, JNTUK, Kakinada, AP, India.

Abstract : This paper proposes an efficient and simple pitch extraction method for Hidden Markov Model (HMM) based speech synthesis. The proposed method uses Dual Tree Complex Wavelet Transform (DTCWT) for pitch extraction. Performance of this method is evaluated using Pitch Tracking Database from Graz University of Technology (PTDB-TUG). Performance evaluations demonstrates that, this method reduces different pitch extraction errors like Voice Decision Error (VDE), Gross Pitch Error (GPE) and F_0 Frame Error (FFE) while extracting pitch from both clean and noisy databases. And also by incorporating this method in HMM based speech synthesis (HTS) system the quality of produced speech is enhanced.

***Index Terms* – Discrete Wavelet Transform, Dual Tree Complex Wavelet Transform, Hidden Markov Model, pitch, speech synthesis, symlets.**

I.INTRODUCTION

HMM based speech synthesis [1]-[2] is a Statistical parametric speech synthesis, is capable of synthesizing speech with various speaking qualities. HMM based speech synthesis can be performed by extracting different speech parameters, that are excitation parameters and spectral parameters. Excitation parameters are the fundamental frequency (F_0) components which in turn give pitch of the signal, and spectral parameters are mel-cepstral coefficients. HMM based speech synthesis can be done in two stages: Training stage and Synthesis stage. In the training stage speech parameters are extracted from speech corpus and modelled as content dependent HMMs. In the synthesis stage speech parameters generate for text input by using the content dependent HMMs. The parameters that in turn used to produce speech signal by using synthesis filter.

Pitch is the quality of a sound governed by the rate of vibrations producing it (or) fundamental frequency (F_0) of a sound wave. Pitch extraction is one of the main aspects in HMM based speech synthesis. As errors in the pitch extraction degrades the performance, different methods [3]-[6] are existed for better pitch extraction. Robust Algorithm for Pitch Tracking (RAPT) algorithm [3] gives erroneous voice decision at low frequencies, and F_0 also varies abruptly. Another Algorithm called Yet Another Alogorithm for Pitch Tracking (YAAPT) is explained in [4] uses original and squared signals for F_0 tracking. This method mainly concentrates on pitch extraction from noise database, but it does not show any enhancement in pitch extraction from clean database. Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum (STRAIGHT) method is explained in [5] uses wavelet-based instantaneous frequency analysis technique. But this method detects faulty in creaky regions. Zero Frequency Filtering (ZFF) method [6], in which pitch is estimated using frame wise ZFF of a signal and appropriate window length. This method gives erroneous F_0 estimation in creaky regions. To avoid detection and estimation faults in creaky regions Continuous Wavelet Transform (CWT) based method [7] is used. It detects creaky regions correctly but it fails in pitch detection using noisy data. Dual Tree Complex Wavelet Transform (DTCWT) [9]-[10] gives better time-frequency representation of the signal and also eliminates noise by successive filtering of the data . So to detect pitch in noisy data and also to improve performance, a method using DTCWT is proposed in this paper.

The rest of the paper is structured as follows: HMM based speech synthesis technique is described in Section.II. The details about Discrete Wavelet Transform (DWT) and Dual Tree Complex Wavelet Transform is given in Section.III. Pitch extraction using proposed method is explained in Section.IV. Section.V describes about the performance evaluations. In section.VI., the concluding remarks are discussed.

II.HMM BASED SPEECH SYNTHESIS

The Hidden Markov model [11]–[13] is one of statistical time series models widely used in various fields like speech synthesis and speech recognition. The HMM based speech synthesis system is shown in Fig.1. The system comprises of training stage and synthesis stage.

In the training stage, by using a speech database content dependent HMMs are trained. The static features from the speech database such as Spectrum and F_0 are extracted at each analysis frame and modeled by multistream HMMs. The output distributions for spectral and $\log F_0$ modelling are continuous probability distribution model and multi space probability distribution model respectively. After modelling, decision tree based context clustering technique is applied to spectral and F_0 models separately.

In the synthesis stage, first, an arbitrarily given text is transformed into a sequence of context-dependent phoneme labels. Based on the label sequence, a sentence HMM is constructed by concatenating context-dependent phoneme HMMs. From the sentence HMM, spectral and F_0 parameter sequences are generated using state duration distributions. Finally, by using an Mel Log Spectral Approximation (MLSA) filter , speech is synthesized from the generated parameter sequences.

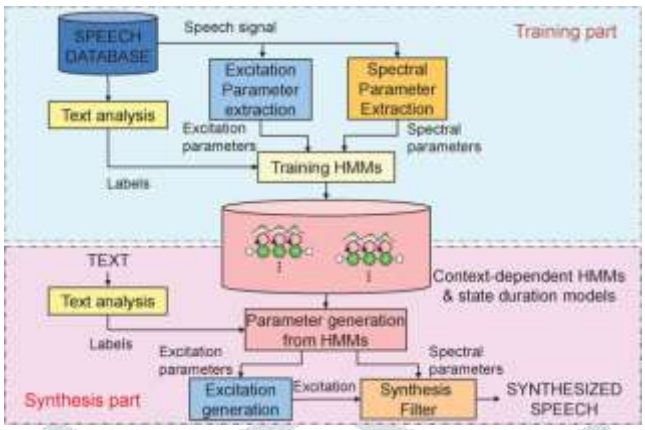


Fig 1. HMM Based Speech Synthesis System

In HTS system, F_0 extraction is done by a pitch extraction technique. Errors present in pitch extraction degrades the performance of the HTS system. So, to reduce the errors in pitch extraction a method is proposed in this paper using DTCWT.

III.DWT & DTCWT

The DWT [8] is computed by successive decomposition of the discrete time-domain signal using low pass filter (g) and high pass filters (h) as shown in Fig.2. At each decomposition level, detail coefficients (*cD*) and approximation coefficients (*cA*) are produced by the high pass filter and low pass filter respectively, followed by a down sampling.

Fig.2. One-Dimensional DWT

The DTCWT [9] is an enhancement to the Discrete Wavelet Transform (DWT) with additional properties like shift-invariant and directionally selective in two and higher dimensions. The DTCWT computes the complex transform of a signal using two separate DWT decompositions (tree a and tree b).as shown in Fig.3.

Proposed method uses Symlets to compute DTCWT coefficients. The symlets are nearly symmetrical wavelets proposed by Daubechies as modifications to the db family. Wave functions of different order Symlets are shown in Fig.4. Since Symlets are more resemble to the speech signal, those are consider in the proposed method. The number of decomposition stages (n) used in the proposed method depends on the sampling frequency (f_s) of the speech signal. As this method extracts frequency in the range of 50-500 Hz, The number of decomposition stages (n) must satisfy the relation(1).

$$\frac{f_s}{2^{n+2}} < 500 \leq \frac{f_s}{2^{n+1}} \tag{1}$$

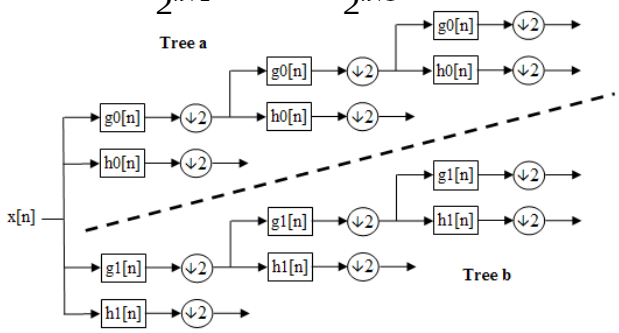


Fig.3.Block diagram for a 3-level DTCWT

speech signal have very low frequency range, the proposed method concentrates mainly on low frequency component at the last stages. The output at the last stage low pass filter eliminates most of the ripples and noise present in the signal. So the required low frequency component only available at the output stage.

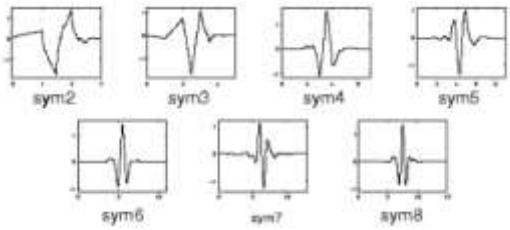


Fig.4. Wave functions of different order Symlets

IV.PROPOSED METHOD

Pitch extraction consists voicing decision and F_0 estimation. Generally, speech contains some voiced signals and non-voiced signals. And only voiced signals have some frequency. So, classification of voiced and unvoiced parts is done first, after that for voiced parts, the corresponding frequency is estimated. In the proposed method, to extract pitch from the speech signal, firstly, speech signal is divided into frames with a frame size of 32ms and frame shift of 10ms. For each frame DTCWT is applied and approximate coefficients at last stage are found. Since this approximate coefficients eliminate the high frequency components, only low frequency components available at the output. Since this low frequency part is the required speech signal, number of peaks at each frame is calculated. Here peaks are present in this low frequency component only when voiced signal is present. Otherwise it shows no variations in the signal. And also, this method calculates peaks in both positive and negative sides and takes the maximum of those two. For voicing decision an approximate threshold value is set to 2. This optimal threshold is found by computing voicing decision errors for different values as shown in Table.1. From the table, it is found that voicing decision is optimum for the threshold of 2. The frames with number of peaks greater than or equal to the threshold are considered as voiced.

Table. 1. VDE for different Thresholds

Threshold	VDE
1	14.5
2	1.22
3	1.76

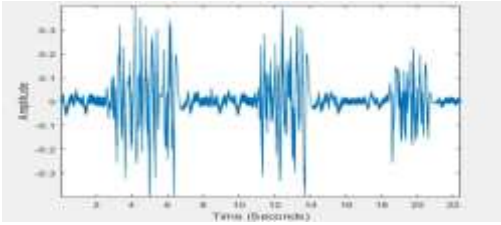
For voiced frames pitch is calculated by using already considered peaks in voicing decision. Since speech signal only available at the output without any secondary's and noise, there is no need of any further processing. Only peak to peak distance calculation is needed. To estimate the F_0, the average distance between the successive peaks is calculated. Since frequency is the reciprocal of time, the inverse of the average distance gives the required F_0. The steps in the proposed method are summarized below:

1. Divide the speech signal into frames, with a frame size 32ms and frame shift 10ms.
2. Compute frame wise DTCWT approximate coefficients.
3. Calculate number of peaks and consider '2' as the optimal threshold value.
4. The frames with number of peaks greater than or equal to threshold is consider as voice frames.
5. For voice frames calculate the average distance between the frames
6. Inverse of voice average distance gives the F_0.

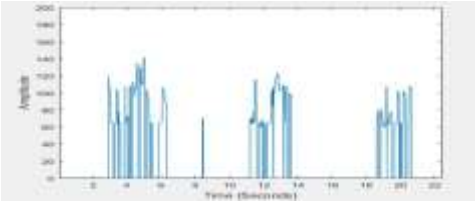
As this method mainly concentrates on low frequency components, it eliminates noise very easily. And also it have very few steps for pitch extraction, which makes it so simple.

V.PERFORMANCE EVALUATION

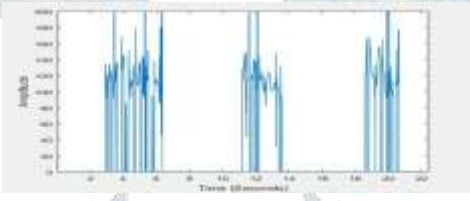
The proposed method is evaluated using PTDB-TUG database. It contain microphone and laryngograph signals of 20 English native speakers as well as the extracted pitch trajectories as reference. This database consists of 4720 recorded sentences totally spoken by both, female and male speakers. For this database corresponding pitch is extracted using YAAPT, CWT based method and proposed method and is shown in Fig.5.



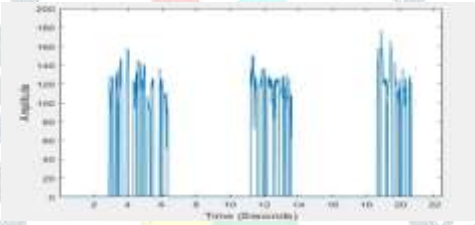
(a)



(b)



(c)



(d)

Fig.5. (a) Original speech signal, its pitch extracted with (b)YAAPT (c)CWT based method and (d) Proposed method

Three measures namely Voicing Decision Error (VDE), Gross Pitch Error (GPE) and F_0 Frame Error (FFE) are used for evaluation.

- **Voicing Decision Error (VDE)** is the percentage of frames for which an error of voicing decision is made.

$$VDE = \frac{F_{error}}{F_{total}} \times 100 \qquad (2)$$

Where *Ferror* = No.of frames in which erroneous Voicing decision made
Ftotal = Total frames

- **Gross Pitch Error (GPE)** is the percentage of voiced frames in which the estimated and the reference pitch differ by more than 20%.

$$GPE = \frac{F_{gp}}{F_{total}} \times 100 \qquad (3)$$

Where *Fgp*= No.of frames with relative error between reference and estimated pitch greater than 20%

- **F_0 Frame Error (FFE)** is the proportion of frames for which an error (either according to the GPE or the VDE criterion) is made. FFE can be seen as a single measure for assessing the overall performance of a pitch tracker.

$$FFE = \frac{F_{error} + F_{gp}}{F_{total}} \times 100 \quad (4)$$

Table.2. and Fig.6. shows the comparison of performance of the three methods (YAAPT, CWT based method and proposed methods). This comparison reveals that, the proposed method gives better results than the existing methods.

Table.2. Performance measures of different methods

Error	YAAPT	CWT Based method	Proposed Method
VDE	1.902	1.76	1.22
GPE	7.06	1.49	1.08
FFE	8.96	3.26	2.3

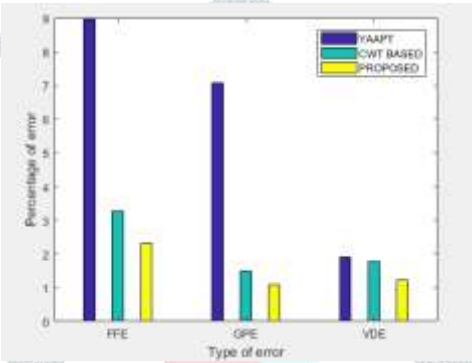
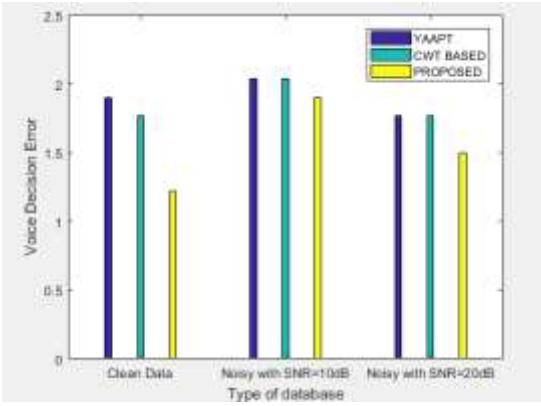


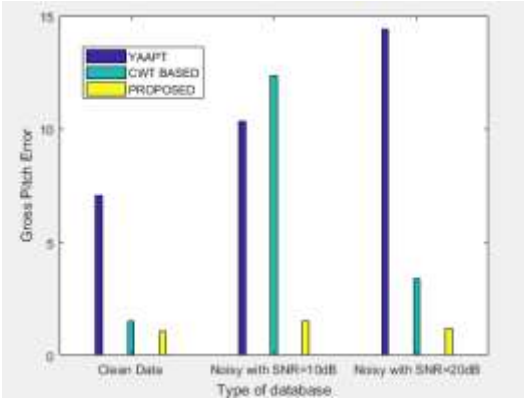
Fig.6. Comparison of Performance Measures for clean database

To evaluate the performance with noisy data a Gaussian noise with Signal to Noise Ratio (SNR) 10dB and 20dB are added to the database. And when this database is used for pitch extraction, the errors occurred due to existing methods are more than errors with the clean database. But proposed method gives almost similar errors as in the case of clean database. This results are shown in Table.3. and Fig.7. From the results, it is observed that voice decision error in noisy database is almost same as voice decision error in clean data for all methods. For noisy database, the gross pitch error and F_0 frame errors are two times and five times as that of clean database in YAAPT method and CWT based method respectively. But, gross pitch error and F_0 frame errors are almost same for both clean and noisy database in proposed method. And also proposed method gives less errors than YAAPT and CWT based methods for noisy database.

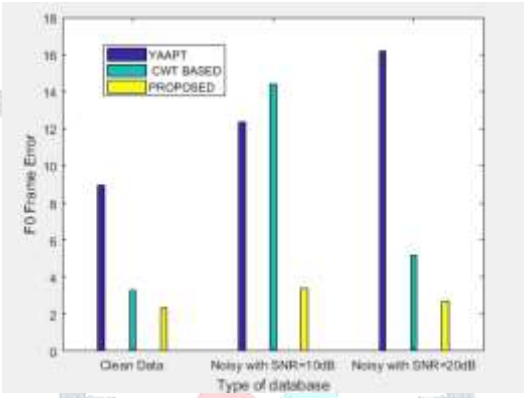
An HTS system is trained with CMU Arctic database. This database comprises 1132 phonetically balanced sentences. This HTS system is built with the Flite HTS engine. The existing (CWT based method) and proposed methods are incorporated in that. Then produced speech is tested with 10 evaluators. The evaluators are youngsters of age group 20-35 years having sufficient speech knowledge. They are asked to give preference between those. Then 70% of evaluators prefer the proposed method. Here due to reduction in the errors of pitch extraction, quality of the synthesized speech is improved.



(a)



(b)



(c)

Fig.7. Comparison of errors for clean and noisy data (a) VDE (b) GPE (c) FFE

Table.3. Performance measures of different methods with clean and noisy database

		YAAPT	CWT Based Method	Proposed Method
VDE	Clean	1.9	1.76	1.22
	SNR with 10dB	2.4	1.9	1.35
	SNR with 20dB	1.6	1.7	1.35
GPE	Clean	7.06	1.49	1.08
	SNR with 10dB	10.3	10.1	1.63
	SNR with 20dB	14.2	2.9	1.2
FFE	Clean	8.9	3.2	2.3
	SNR with 10dB	12.7	12.09	2.9
	SNR with 20dB	15.8	4.7	2.5

VI.CONCLUSION AND FUTURE WORK

This paper proposes a simple and efficient pitch extraction method using DTCWT. By considering the low frequency coefficients of the signal from DTCWT it eliminates high frequency ripples and noise present in the data. Hence, it works better for pitch extraction from both clean and noisy data by reducing different errors like VDE, GPE and FFE. And also due to reduction in the errors in pitch extraction, it improves the quality of speech synthesized, when this method is incorporated in HTS system. For noisy database, this method works better for only SNR value greater or equal to 10. For noisy database with SNR less than 10 this method doesn't give better performance. So, in future, it can be developed for noisy data with SNR less than 10.

References

[1] J. Yamagishi, "An introduction to HMM-based speech synthesis," Tokyo Inst. Technol., Tech. Rep., 2006.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis," in *Proc. Eurospeech*, Budapest, Hungary, vol. 5,pp. 2347–2350, 1999.

[3] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis*. Amsterdam, Netherlands: Elsevier Science, 1995, ch. 14.

[4] Kavita Kasi, Zahorian S.A., "Yet Another Algorithm for Pitch Tracking" in International Conference on Acoustics ,Speech, and Signal Processing, 2002.

[5] H. Kawahara, H. Katayose, A. de Cheveigne, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurateestimation of f0 and periodicity," in *Proc. Eurospeech*, vol. 6, pp. 2781–2784, 1999.

[6] N. P. Narendra and K. Sreenivasa Rao, "Robust voicing detection and *F*0 estimation for HMM-based speech synthesis," *Circuits, Syst., Signal Process.*, vol. 34, no. 8, pp. 2597–2619, 2015.

[7] M. Kiran Reddy, K. Sreenivasa Rao, "Robust Pitch Extraction Method for the HMM-Based Speech Synthesis System", IEEE signal processing letters, vol. 24, no. 8, 2017

[8] I. Daubechies, *Ten Lectures On Wavelets*. Piladelphia, PA: SIAM, 1992.

[9] N.G. Kingsbury, "The dual-tree complex wavelet transform: A new efficient tool for image restoration and enhancement," in *Proc. European Signal Processing Conf.*, Rhodes, Sept. 1998, pp. 319–322.

[10] N.G. Kingsbury, "The dual-tree complex wavelet transform: A new technique for shift invariance and directional filters," in *Proc. 8th IEEE DSP Workshop*, Utah, Aug. 9–12, 1998, paper no. 86.

[11] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov models for speech recognition*, Edinburgh University Press, 1990.

[12] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, Prentice-Hall, Englewood Cliffs, N. J., 1993.

[13] S. Young, G. Everman, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book Version 3.2.1*,December 2002.

[14] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[15] "HMM-based speech synthesis system (HTS)." [online]. Available: http://hts.sp.nitech.ac.jp/

[16] "ESPS software package." [online]. Available: http://www.speech.kth.se/ software/#esps.

[17] S. Mallat, *A Wavelet Tour of Signal Processing*. New York, NY, USA:Academic, 1999.