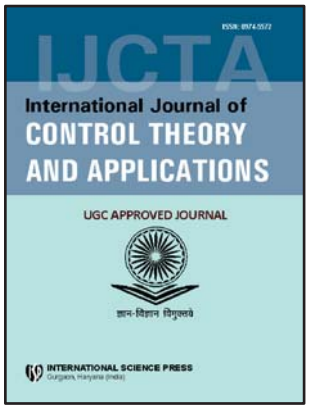

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 32 • 2017

VHDL Implementation of Voice Morphing Features in terms of Magnitude and Angle using FFT

Shaikh S.A.[a] and Godbole B.B.[b]

[a]*Research Scholar Dr. B.A.M.U. Aurangabad, India.*

E-mail: shaikhshoaibarif@gmail.com

[b]*Associate Professor of KBPCEP. Satara Shivaji University, India*

E-mail: bbgodbole@rediffmail.com

Abstract: A less complex structure of voice morphing to support VLSI feasibility with lesser cost in terms area, and power is developed in this paper. There are two signals taken separately as a source and target. Both signals are transform time domain to frequency domain with help of 3 stage 8 point FFT.Thetransformed information gives the amplitude andphase value, further these values are adjusted with FFT shift.The IP is implemented in VHDL and synthesized in Xilinx Spartan 6(6*slx* 16*csg* 324-2) device. The design is thoroughly tested in Modelsimsimulation.

Keywords: *FFT, IP, RAM, VHDL, Voice Morphing.*

1. INTRODUCTION

When the voice communication occurs in the recording of digital video and audio broadcasts network, oral replies to the questions of various interviews, the speaker often required to preserve their anonymity. In such instances, the means of cover and change voice - morphing technology.

A lot of software were created to change the human voice, working in real time, and with the registered audio recordings. A very popular change of voice tone by raising or lowering the pitch of his/her at random.

However, when using professional tools identify these changes cannot guarantee security. Also, people who are personally acquainted with the speaker, can identify his manner of speech, because urgent task is to supplement the new features morphing technology.

The aim of this work is to develop a morphing algorithm in VHDL with new properties that will reduce the opportunity to learn even speaking to people personally familiar with it, and its implementation. The main objective of the work - the creation of an algorithm that modifies the voice in such a way for which it is difficult or even impossible to identify the voice of the speaker. To do this, it needs to examine the basic techniques of morphing and choose among them a suitable, supplementing it with new features, or create your own new algorithm.

1.1. Physiology of Speech Production

The acoustic speech signal is the result of coordinated movements of the human vocal apparatus. The movement of air in the vocal tract (Figure 1) is provided by the pressure caused by the respiratory muscles of the lungs. The very vocal tract is represented by the larynx and air cavities, the configuration of which is changed in the process of speech production.

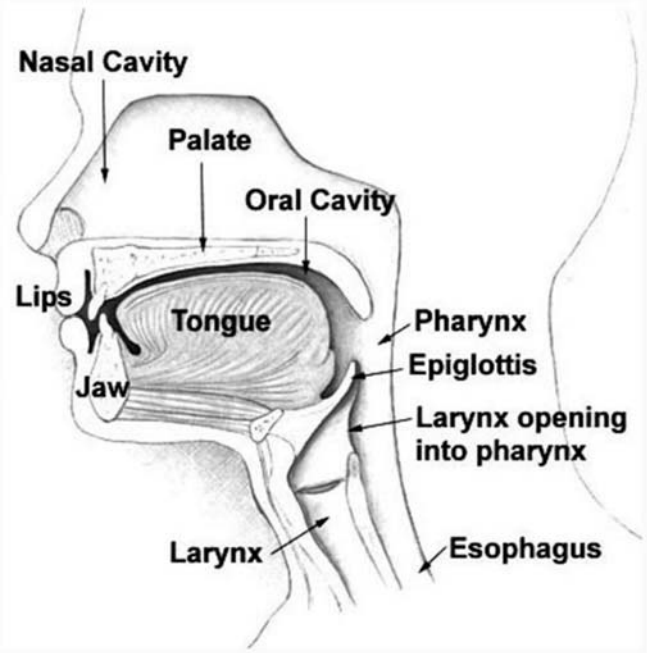

Figure 1: Human vocal tract

Organs involved in the formation of the speech can be divided into active (movable in the formation of sounds) and passive (supporting) organs. Among the active organs of speech: vocal cords, tongue, lips, soft palate, uvula, the back throat and lower jaw. Passive organs of speech and define the shape of the resonance properties of the cavities. The passive organs include the teeth, alveoli, solid palate and upper jaw [1].

The jet of air exiting from the trachea, must pass through the cricoid and the gap between the vocal cords, the position of which depends on whether there is breath, whispers or uttered sounds of a certain height[2].

The vocal cords are a valve mechanism, which allows vibrations to obtain sounds of different frequencies - the basic tone of voice. Next, the sound system passes air through the vocal tract areas in which there is sound frequency filtering [3]. The decisive role in the formation of speech sounds played by the movement of the soft palate, tongue, lips and lower jaw. The nasal cavity is the cavity, increasing the vibrations of certain frequencies [4].

1.2. Speech Synthesis Methods

Voice morphing technology is, first and foremost, a tool of artificial speech synthesis. For speech synthesis, there are different approaches.

1. The method of encoding-recovery waveforms;
2. Digital modeling of the vocal tract.
3. Analog synthesis method of formant frequencies.

The first approach is the most basic and simple, involves a concatenation of pre-recorded sound samples. Phrases and words are recorded separately and played back at the right time for the appropriate program commands. This method is suitable in cases where the synthesizer must be legible to say a relatively small and well-known pre-set phrases, nonetheless drawn up at the junction of sound fragments intonation possible distortions and gaps, marked by ear. To play a custom text created by the base, where the units are phonemes and allophones, but taking into account all the peculiarities of pronunciation require large memory resources and labor costs [1].

Digital modeling of the vocal tract method also involves the creation of a dictionary with the participation of speaker: no written words and phrases, and the frequency and voice speech options that can reduce the amount of memory required for speech output.

The synthesis method of formant frequencies is not based narration. It is based on the phonetic dismemberment of speech on the minimum identifiable unit of time - phonemes - and played back by hesitation repeating their formant construction. Formants in this case can be called analogues phonemes in the frequency scale.

The advantage of this method is its versatility, as it is created separately from the spoken sound, which makes it possible to have an unlimited vocabulary. Among the shortcomings - promiscuity speech synthesizer and the difficulty in complying with the rules of pronunciation specific to language. Synthesizer forms a sequence of phonemes by which copies the natural human speech. Formant synthesis method is sometimes described as the digital modeling of the vocal tract.

Various synthesis areas are considered and in most concatenation of speech sounds: popular Diphone and allophones approaches and Unit Selection technology [5]. Methods for Diphone synthesis and allophones are generally similar, and differ only in the audio base units (using the boundaries of allophones units are less visible, as are natural places spectral changes). In the case of Unit Selection speaker speech segmentation techniques to sound produced based on various levels so as to be smooth areas of the signal of longer duration, but due to the limited size of some voice base elements may be distorted or even to fall out of the question. In order to achieve a compromise between sound quality and the used memory resources can be used hybrid speech synthesis, combining technology and Unit Selection allophone synthesis [6].

Various techniques for voice conversion have been developed during last two decades [7-12]. Out of those the very first approach is linear predictive coding (LPC) [13]. In LPC, the excitation signals are generated using the concept of residual error [7, 9 and 10]. A lot of research work is focused on following criteria:

1. Linear Prediction Coding (LPC).
2. Line Spectral Frequencies (LSFs) [14].
3. Harmonic plusnoise model parameters [15].
4. Interpolation of speech parameters and modelling the speech signals using format frequencies [8],
5. Mixed timeand frequencydomain methods to alter the pitch, duration, and spectral features.

Aforementioned techniques exist under singlescale morphing.

Statistical speech representations [16] may be transformed to target using various speaker adaptation methods initially established for speech recognition [17].Because of short window and non-stationary behavior of speech signals a rapid adaptation is required as factorization techniques like Eigen-voices [18]. Non-linear methods have also been offered in the form of kernel regression [19].Recent works in improvement of quality of speech to predict to predict the acoustic parameters from the linguistic ones [20] and Vocaine as the vocoder[21]. Neural network based approach has been also consider efficiently in source target adaptation[22. 23].

Vocal tract resonances, log area vocal tract functions are considered in voice morphing [24]. Wavelet based pitch shifting technique has been proposed in [24] and pitch is measured using autocorrelation in wavelet domain.

This paper presents a FFT based algorithm which is hardware friendly and simultaneously and optimally estimates the frequency warping and the frequency weighting. Section 2 presents proposed methodology used in this research work. Section 3 shows simulation results and finally the conclusion is presented in Section 4.

2. PROPOSED METHODOLOGY

Voice transformation is an operation that involves modifying the recordings audio in order to change the perceived identity. For example, if one wants to create a voice recording of a man from a woman'svoice, change the timbre, intonation, emphasis or pronunciation on certain regions, etc.

The objective of this paper consists of carrying out comparative analyses on different pairs of voices (the source and the target) pronouncing the same text, in order to discern the most important differences from a perceptual point of view:

The timbre and the prosody (the height, the energy and the duration of the phones and the silences). During this training step, therefore, locally compares audio extracts corresponding to the same phonemes pronounced but present in the signal under the form of different acoustic events.

One of the main problems is the difficulty of perfectly aligning the voices source and target without knowing a priori their respective behaviours in a context of precise pronunciation.

As an objective to imitate a precise voice (called target) by transforming the recordings of a source voice. This paper creates the perceptual illusion that the target has pronounced what the source has recorded.

A simplified flowchart showingthe procedure used here to achieve speech morphing isillustrated in Figure 2.

A VHDL implementation of Voice morphing IP according to Figure 2 contains RAM, FFT and Pitch calculation blocks majorly. There are source and target signals, can be carried out by altering. So as per the paper, at first samples of source and destination voice are converted to frequency domain using FFT. Then difference between both is calculated, it gives info that how source is different from target. Then full source voice is varied by that difference thus we get morphed voice to target.

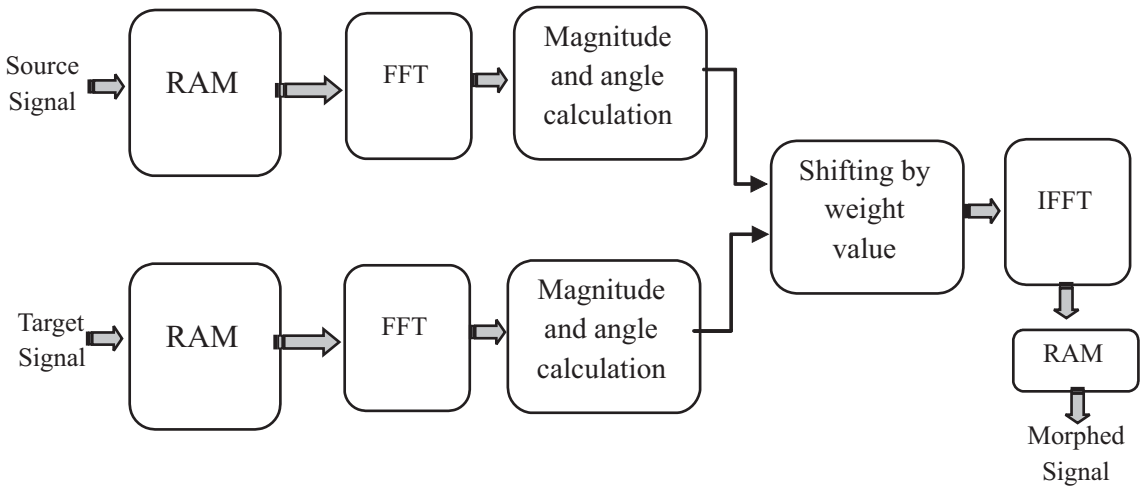

Figure 2: Flowchart of the speech morphing algorithm

RAM : It is an 8 bit read only memory, which takes the data input form external resource. There are two separate input source and target, which is further read by FFT block. Once entire process of voice morphing is over then morphed signal are saved in RAM.

FFT : The most well-known and used algorithms are the FFT algorithms where N is an integer power of two, N = 2M, where M is an integer. It is possible toreduce the number of operations necessary to an order of magnitude of N log2(N) = N × M

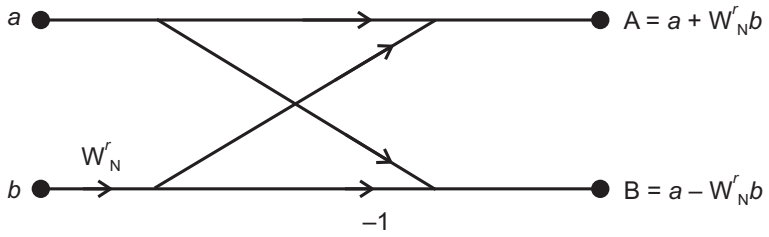

Figure 3: Butterfly structure

The Radix-2 based FFT algorithm is one of the algorithms with the butterfly structure which is the simplest for the calculation of the FFT. It is an 8 bit 3 stage 8 point FFT.The butterfly structure is implemented in Figure 3:

$$W_N^{k+n/2} = e^{-j\left(\frac{2\pi k}{N}+\pi\right)}$$

$$= e^{-\frac{j2\pi k}{N}}$$

$$= -W_N^k \quad (1)$$

Therefore,
$$W^{(N)}(k) = X_0^{\left(\frac{N}{2}\right)}(k) - W_N^k X_1(k) \quad (2)$$

for
$$k = 0, \ldots, \frac{N}{2} - 1$$

$$X^{(N)}\left(k+\frac{N}{2}\right) = X_0^{\left(\frac{N}{2}\right)}(k) - W_N^k X_1(k) \quad (3)$$

for
$$k = 0, \ldots, \frac{N}{2} - 1$$

$$X(0) = \sum_{n=0}^{0} x(n) e^{-j\left(\frac{2\pi.0}{2}\right)n} \quad (4)$$

Amplitude
$$|X[k]| = \sqrt{x_{re}^2 + x_{im}^2} \quad (5)$$

Where x_{re} and x_{im} represent real and imaginary parts respectively.

Phase angle
$$\angle X[k] = \tan^{-1}\left(\frac{x_{im}}{x_{re}}\right) \quad (6)$$

With optimal change of the pitch of speech signals and match the nearest shift of target speech signal is the key factor of morphing. To do the match source calculated amplitude and angle is shifted with certain weight value to match with target phase and amplitude. This interpolation creates an intermediate sound in the morph.

3. SIMULATION AND RESULTS

3.1. RTL Schematics

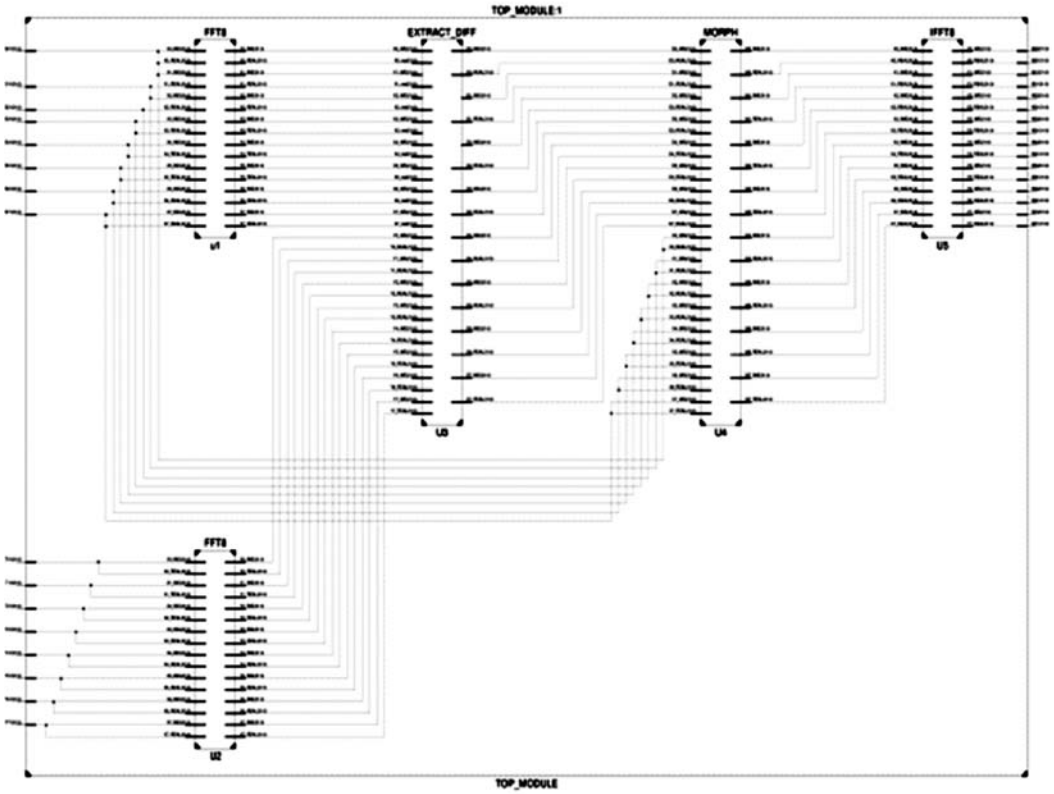

Figure 4: RTL schematic for top entity

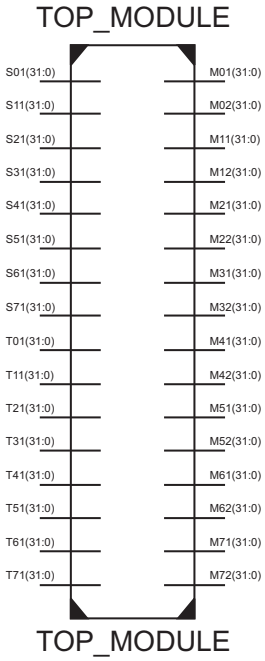

Figure 5: Top entity

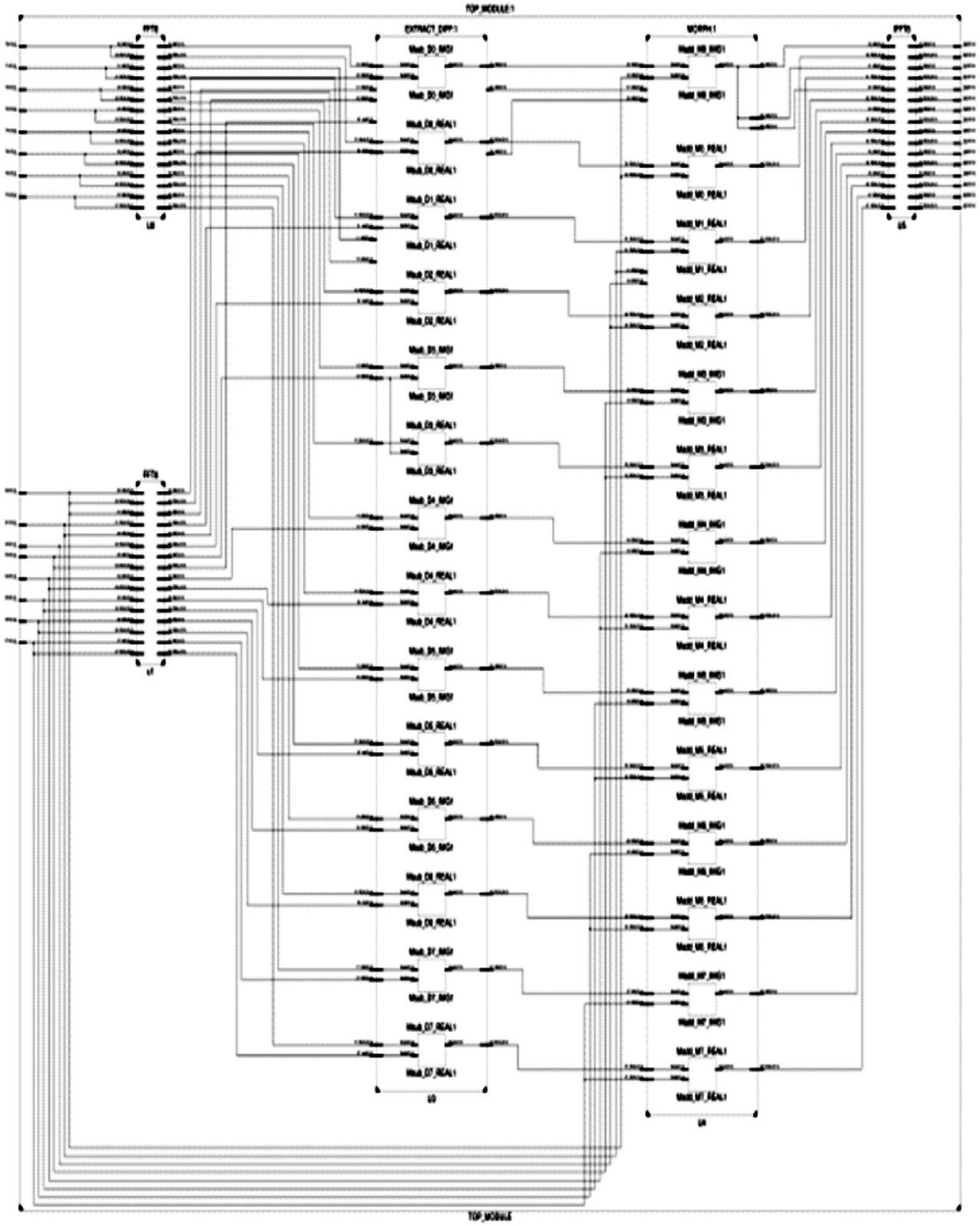

Figure 6: RTL schematic for morphing block

3.2. Simulation Waveform

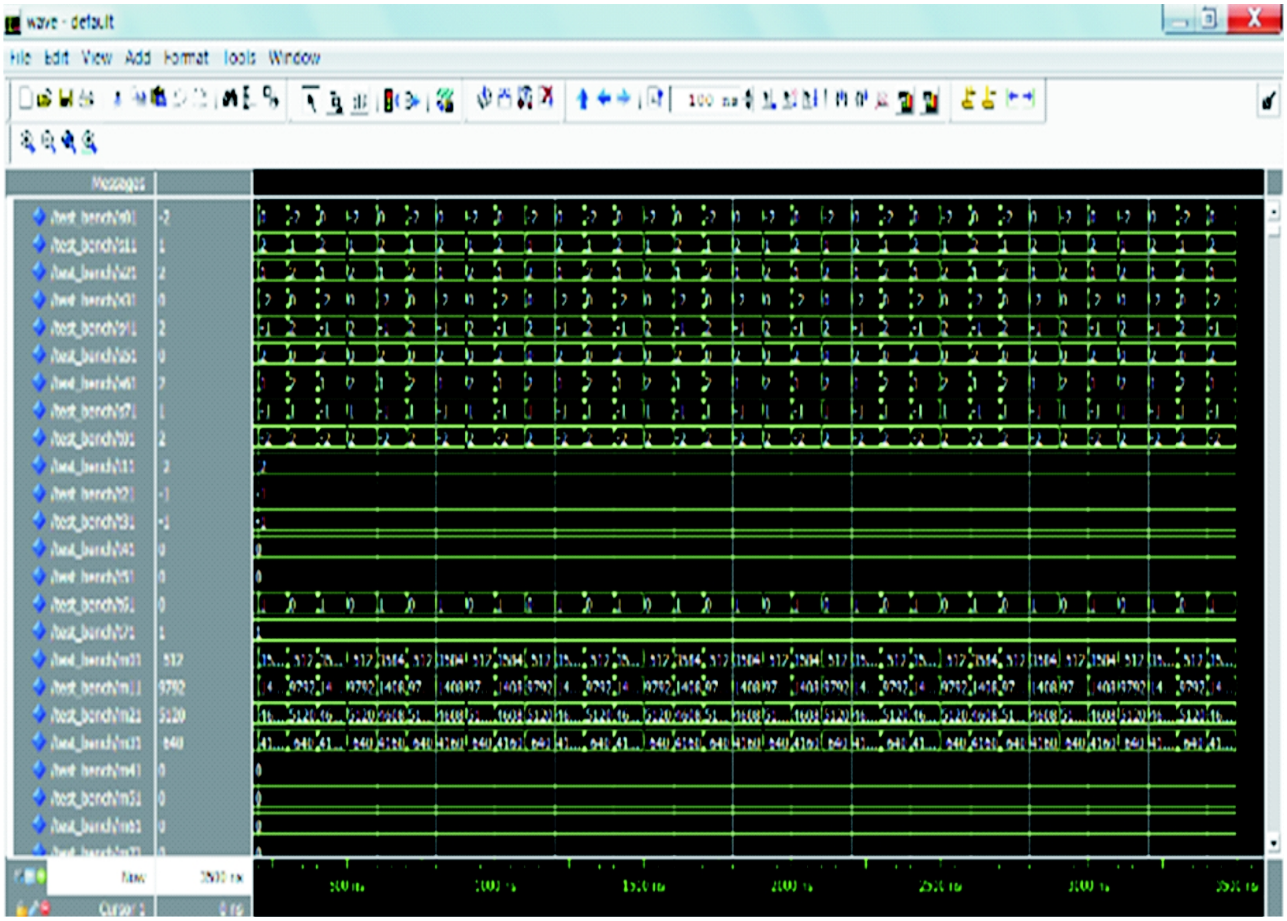

Figure 7: Simulation waveform for voice morphing

Device utilization summary:

Selected Device: **6slx16csg324-2**

Slice Logic Utilization:

Number of Slice LUTs: 7535 out of 9112 82%

Number used as Logic: 7535 out of 9112 82%

Slice Logic Distribution:

Number of LUT Flip Flop pairs used: 7535

Number with an unused Flip Flop: 7535 out of 7535 100%

IO Utilization:

Number of IOs: 1024

Number of bonded IOBs: 1024 out of 232 441% (*)

Specific Feature Utilization:

Number of DSP48A1s: 16 out of 32 50%

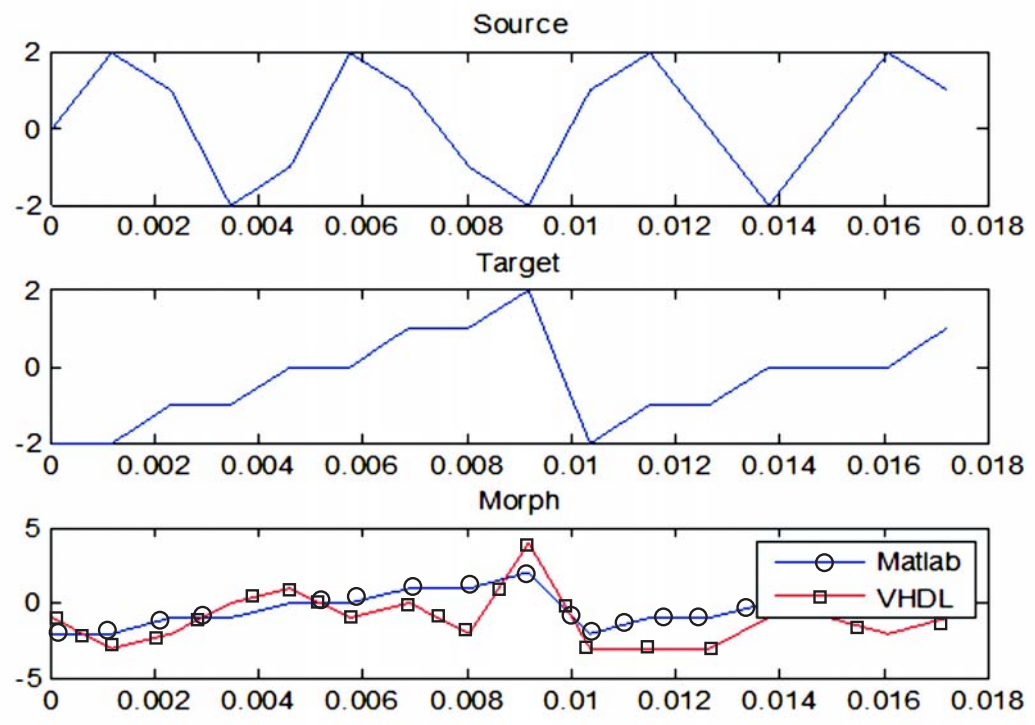

Figure 8: Comparative analysis for same prototype implemented in MATLAB and VHDL

Figure 8: shows the separate plot of source and target according to their time and amplitude. Source resembles the sinusoidal nature at the same time in target resembles random nature of wave. At .002-time source amplitude is shifted to match the nature of target wave. Simulation analysis is done in MATLAB and Modelsim. A VHDL implementation is tested with certain test-bench by giving some values of speech signals obtained from MATLAB interface. This signals is further processed with FFT (butterfly structure) and stretched outcome is stored in RAM. The same results is stored in 'output.txt' with 'textio' option in VHDL, and stored as text file. This morphed 'textfile' is further read in MATLAB. Comparative analysis shows the similar nature of outcome form VHDL IP v/s MATLAB.

4. CONCLUSION

The IP of voice morphing is developed according to hardware requirement. Keeping the objective of less area and less power robust structure is proposed. The design is thoroughly tested by simulation .There is separate prototype was built in MATLAB with existing keywords of FFT and when it is compared with VHDL implementation of prototype. It is giving near about similar results; hence simulation concludes the preciseness of implementation. After FFT there is complex output, we have not included the complex floating point architecture, in this paper, in future floating point structure can be considered to get the more precise results. The IP is implemented in Spartan 6 (6slx16csg324-2) device.

REFERENCES

[1] Keller, E. (1995). Fundamentals of speech synthesis and speech recognition: basic concepts, state-of-the-art and future challenges. John Wiley and Sons Ltd.

[2] Wardhaugh, R. (1972). Introduction to Linguistics. Birkholz, P., & Kröger, B. J. (2007, August). Simulation of vocal tract growth for articulatory speech synthesis. In Proc. 16th Internat. Congress of Phonetic Sciences (Saarbrücken, Germany) (pp. 377-380).

[3] Lieberman, P., & Blumstein, S. E. (1988). Speech physiology, speech perception, and acoustic phonetics. Cambridge University Press.

[4] Fant, G. (1971). Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations (Vol. 2). Walter de Gruyter.

[5] Rutten, P., Aylett, M. P., Fackrell, J., & Taylor, P. (2002, September). A statistically motivated database pruning technique for unit selection synthesis. In INTERSPEECH.

[6] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013). Speech synthesis based on hidden Markov models. Proceedings of the IEEE, 101(5), 1234-1252.

[7] L.M. Arslan, D.Talkin,"Voice conversion by codebook map-ping of line spectral frequencies and excitation spectrum," Proc. Eurospeech, pp.1347-1350, 1997.

[8] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara: Voice conversion throughvector quantization. IEEE Proceedings of the IEEE ICASSP, 1998, 565–568.

[9] L. Arslan: Speaker transformation algorithm using segmental codebooks (stasc).Speech Communication 28, 1999, 211–226.

[10] Y. Stylianou, O. Cappe, and E. Moulines: Statistical methods for voice quality transformation. Proc. EUROSPEECH, 1995, 447–450.

[11] http://www.seminarpaper.com/2011/12/voice-morphing-full-report.html.

[12] Z. Shuang, F. Meng, Y. Qin, "Voice Conversion by Combining Frequency Warping with Unit Selection", in Proc.ICASSP, pp.4661-4664, 2008.

[13] Bradbury, J. (2000). Linear predictive coding. Mc G. Hill.

[14] A. Kain and M.W.Macon: Spectral voice conversion for text-to-speech synthesis.Proc. ICASSP'98 1, 1998.

[15] H. Valbret, E. Moulines, and J.P. Tubach: Voice transformation using PSOLA technique. Speech Communication 11, 1992, 175–187.

[16] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Review: Statistical parametric speech synthesis," Speech Communication, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[17] Junichi Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech syn- thesis and a constrained smaplr adaptation algorithm," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 1, pp. 66–83, Jan 2009.

[18] Tomoki Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to- one voice conversion based on eigenvoices," in ICASSP, April 2007, vol. 4, pp. IV–1249–IV–1252.

[19] Hanna Silen, Jani Nurminen, Elina Helander, and Moncef Gabbouj, "Voice conversion for non-parallel datasets using dynamic kernel par- tial least squares regression.," in Interspeech. 2013, pp. 373–377, ISCA.

[20] Heiga Zen and Hasim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in ICASSP, 2015, pp. 4470–4474.

[21] Yannis Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in ICASSP, 2015.

[22] Orhan Karaali, Gerald Corrigan, and Ira A. Gerson, "Speech synthesis with neural networks," CoRR, vol. cs.NE/9811031, 1998.

[23] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in ICASSP. IEEE, 2013, pp. 7962–7966.

[24] Kumar, A., & Jain, R. (2006, September). Speech pitch shifting using complex continuous wavelet transform. In 2006 Annual IEEE India Conference (pp. 1-4). IEEE.