# Time and Pitch Scale Modification of Audio Signals using Short Time Fourier Transform

Srinivas Rao Chintagunta[1] Mrutyunjaya Nanda[2] Rajib Lochan Swain[3]
Asst. Prof., Department of Electronics and Telecommunication, IIIT Bhubaneswar, India
Department of Electrical and Electronics Engineering, IIIT Bhubaneswar, India
Department of Electrical and Electronics Engineering, IIIT Bhubaneswar, India

## ABSTRACT

A method for independently modifying the time and pitch scale of acoustic signals, with an emphasis on speech signals is proposed in this paper. The algorithm developed here is based on Short Time Fourier Transform(STFT).The purpose of this paper is to devise a way to change the rate of a pre-recorded sound without altering the frequency content. Simply playing the sound at a different rate is not a solution. The frequencies would be distorted in proportion to the scaling factor, and at very low or high rates, would be very difficult to understand at all, let alone identify as human speech. Our approach is to sample the digital signal and then interpolate data points between our samples to produce a sound of the desired length. A slowed-down sound would have more points inserted between the samples than the original signal had, and a speeded-up sound would have fewer than the original. Performance of the proposed algorithm is demonstrated using spectrum plots.

**Keywords**—Speech analysis, speech processing, time scale modification, wavelet packet transform

## 1. INTRODUCTION

There are many important uses for slow-down speech or music. Similarly, both music lovers and musicians would benefit from the ability to slow down music, which will slowdown the recording themselves and improve their performance. We need to slow down the signal during foreign language translations, or for hearing-impaired listeners. In other applications it is also useful to be able to increase the rate of articulation, so that the material may be scanned quickly. Also useful is the ability to increase the speed of speech for commercials, both television and radio. Because time is so valuable, being able to communicate more information about their product in a given amount of time would be very valuable to advertising agencies and their clients. In other applications, the speech is compressed or expanded in frequency. In particular, frequency compression is useful in bandwidth reduction. There various approaches used earlier such as time domain splicing/overlap-add [3] ,[6] ,[7] are computationally cheap, but suffers from echoes. Frequency domain approaches [4],[5],[8] are computationally intensive, and provides high quality output.

To accomplish the requirement of high-quality time scaling of speech signals, a number of algorithms have been proposed in the past decade. Unfortunately, applying these algorithms on music signals does not yield satisfactory results. The proposed algorithm is related to the PSOLA-like algorithms, which are based on the Short-Time Fourier Transform between the original and the time-scaled signal. The basic assumption of these algorithms is that the spectral characteristics of the signal are constant for short durations, and that the signal is quasi-periodic in the time domain. The signal in PSOLA is divided into short-time overlapping frames, which are used for constructing the time-scaled synthesis signal, while maintaining the original spectral parameters and their related location.

Experimental results show that the proposed model can achieve better performance, in terms of spectral and phase characteristics and the synthetic speech quality, than conventional model in the case of time and pitch scaling of signals. The remainder of this paper is organized as follows. In Section 2, a detailed description of the sinusoidal model is given, Section 3 gives a brief description about time and pitch scale transformation, Section 4experimental results are presented and Section 5 concludes the paper.

## 2. STFT MODEL

The sinusoidal model [1] represents a speech signal as a linear combination of sinusoids with time-varying amplitudes, frequencies, and phases. That is, the speech signal is represented as the sum of a finite number of corresponding sinusoidal parameters at the fundamental frequency. The input speech signal is given by

$$e(t) = \sum_{k=1}^{N} A_k(t) \cos\left(\int_0^t \omega_k(t)dt + \Phi_k\right) \quad (1)$$

Where, N is the number of sinusoids, $A_k, \omega_k and \Phi_k$ are the time varying amplitude, frequency and phase respectively. This approach is very simple and shown in Fig.1. First, we take the STFT of an input signal for that a signal is windowed at overlapping intervals. For each window of data (overlapping frames of audio), the discrete Fourier transform (DFT) is computed. The locations of the peaks of the DFT magnitude function are used as estimates of the frequencies of the underlying sine-wave components. The magnitude and phase of the Fourier Transform at these measured frequencies are used as estimators, of $A_k and \Phi_k$ respectively. The block diagram of STFT model is shown in Fig.1.

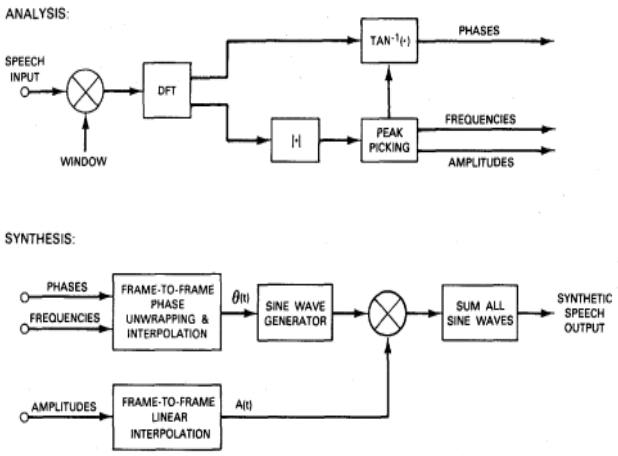


**Fig 1**. Block diagram of STFT model

The reconstructed signal should be

$$s_R^m(t) = \sum_{k=1}^{N(m)} A_k^m(t)\cos(\phi_k^m(t) + \Phi_k^m)$$

$$\phi_k^m(t) = \int_0^t \omega_k^m(t)dt \qquad (2)$$

In order to achieve (2), there are some constraints [1] to interpolating the amplitude, frequency and phase.

## 3. TIME AND PITCH SCALING

The time and pitch scaling algorithms are used to change the speech rate without affecting the pitch or timbre of the speaker voice, or vice-versa – to change the pitch while leaving the rate and the timbre unaffected. The mathematical descriptions of the terms are given in this section**.**

### 3.1. Time Scaling

The process of time scaling a sinusoidally modeled signal by a factor $\rho$ involves the length $T_A$ of analysis frame and length $T_R$ of reconstruction frame to be related as

$$T_R = T_A$$

This implies that the amplitude and frequency information at time *t* should be mapped to new time $t' = \rho t$, so that the scaled signal may be represented as

$$s_R^m(t') = \sum_{k=1}^{N(m)} A_k^m\left(\frac{t'}{\rho}\right)\cos\left(\rho\phi_k^m\left(\frac{t'}{\rho}\right) + \Phi_k^m\right) \qquad (3)$$

This function has been derived over one (the m[th]) reconstruction frame only, and the polynomial phase function $\phi_k^m(t)$ is only valid for $0 < t < T_A$, where $T_A$ is the analysis frame length. The phase term has been multiplied by ρ in order to preserve instantaneous frequency (the derivative of phase) during time scaling.

Of course in practice one would wish to reconstruct, in a time-scale modified manner, a continuous stream of data. This could be achieved most directly by calculating a scaled reconstruction frame for each analysis frame and then simply concatenating successive output frames. Unfortunately, this will result in a strongly degraded result because the time scaling incurs a lack of matching of phases between ssuccessive frames[1].

$$s_R^m(t') = \sum_{k=1}^{N(m)} A_k^m\left(\frac{t'}{\rho}\right)\cos\left(\rho\phi_k^m\left(\frac{t'}{\rho}\right) + \Phi_k^m + \gamma_k^m\right)$$

(4)

Where the phase offset $\gamma_k^m$ is calculated to eliminate the discontinuity by setting it according to

$$\gamma_j^{m+1} = \rho\phi_k^m(T_A) + \Phi_k^m - \Phi_j^{m+1} \qquad (5)$$

Unfortunately, the reconstruction (eq. 4) with the $\gamma_k^m$ offset (designed to preserve phase continuity) destroys the phase information $\Phi_k^m$ in the reconstructed signal when $\rho \neq 1$.

### 3.2. Pitch Scaling

The process of pitch scaling involves every frequency track is scaled by same constant amount say $\sigma$. The pitch scaled signal can be represented as follows.

$$s_p^m(t) = \sum_{k=1}^{N(m)} A_k^m(t)\cos\left(\sigma\phi_k^m(t) + \Phi_k^m\right) \qquad (6)$$

If frames are simply concatenated together, phase discontinuities will occur, so the reconstructed pitch-scaled signal is actually formed as

$$s_p^m(t) = \sum_{k=1}^{N(m)} A_k^m(t)\cos\left(\sigma\phi_k^m(t) + \Phi_k^m\right) + \gamma_k^m \qquad (7)$$

Where the phase offset $\gamma_k^m$ is calculated as

$$\gamma_j^{m+1} = \sigma\phi_k^m(T_A) + \Phi_k^m - \Phi_j^{m+1} \qquad (8)$$

## 4. EXPERIMENTAL RESULTS

In this section, we presented time scaling and pitch scaling performance on 3 seconds duration, 8 KHz sampled record of the speech (IIIT-Bhubaneswar) of male and female voice.

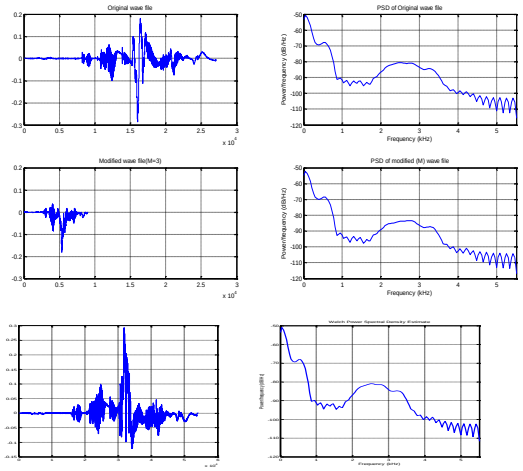

**Fig .4.1:** Time scale modification of speech (female).(a) original (b) compression (c) expansion with corresponding psd.

Fig 4.1., shows female and Fig 4.2., shows male for that (a) original voice, (b) compression with ρ=3 and (c) expansion with ρ=2. In time scaling the time domain speech signal is altered with horizontal (time) axis but the spectrums are same in all these three. The corresponding psds are given at the right side.

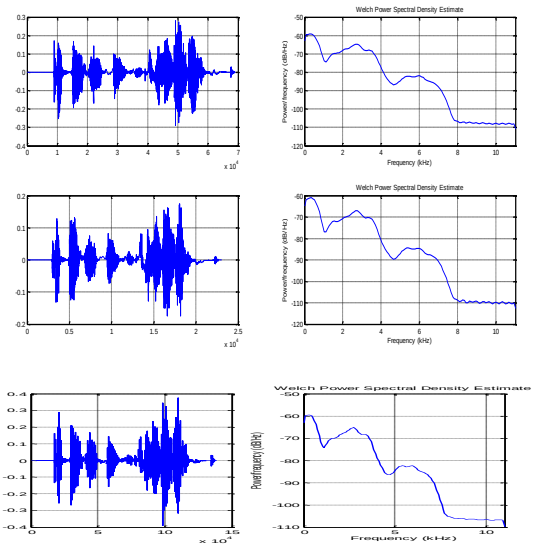

**Fig. 4.2:** Time scale modification of speech (male).(a) original (b) compression (c) expansion with corresponding psd

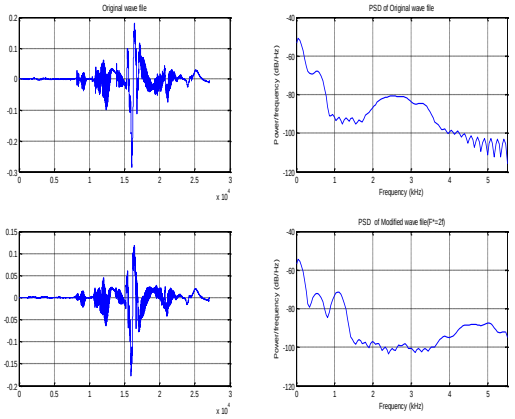

**Fig.4.3:** Pitch modification of speech (female) (a) original (b) Pitch scaled with corresponding psd's.

Fig 4.3., shows female and Fig 4.4., shows male for that (a) original voice, (b) pitch scaled with σ =2. In pitch scaling the time domain speech signals are same but the spectrums are altered. The corresponding psds are given at the right side.

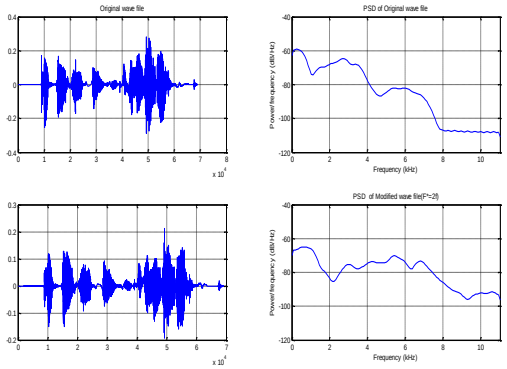

**Fig.4.4:** Pitch modification of speech (female) (a) original (b) Pitch scaled with corresponding psd's.

## 5. CONCLUSION

All the above results are tested of the speech signal by the word IIIT-Bhubaneswar. The performance of STFT for time scaling and pitch scaling of audio signals are discussed here. The scaling method uses parametric modeling techniques to achieve independent time and pitch scaling of audio signals. In time scaling the speech is compressed or expanded without any change in the spectrum. Similarly, in pitch scaling the spectrums are changed without any change in time domain signal. By addressing the phase dispersion defect in preexisting frequency domain based scaling methods it provides better quality transformations. This method also has a computational advantage as it does not require the decomposition of the signal into excitation and vocal tract responses.

## 6. REFERENCES

[1] Brett Ninness and Soren John Henriksen," Time-Scale Modification of Speech Signals" IEEE Trans. on signal Processing, vol. 56, no. 4, Aril. 2008.

[2] E. Hardam, "High quality time scale modification of speech signals using fast synchronised-overlap-add

algorithms," in Proc. IEEE Int Conf. Acust., Speech Signal Process., 1990, pp. 409–412.

[4] R. McAulay and T. Quatieri, "Speech transformations based on a sinusoidal representation," IEEE Trans. Acoust., Speech., Signal Process., vol. ASSP-34, no. 6, pp. 1449–1464, Dec. 1986.

[5] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," Speech Commun., vol.16, pp. 175–205, 1995

[6] W. Verhelst and M. Roelands, "An Overlap-Add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process.1993, pp. 554–557.

[7] J.Wayman, R. E. Reinke, and D.Wilson, "High quality speech expansion,compression, and noise filtering using the SOLA method of time scale modification," in Proc. IEEE Int. Conf. Acoust., Speech SignalProcess., 1989, pp. 714–717.

[8] D. W. Griffin and J. S. Lim, "Signal estimation from modified shorttime Fourier transform," IEEE Trans. Acoust., Speech, Signal Process.,vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[9] Robert J. McAulay and Thomas F. Quatieri, *Speech Analysis/Synthesis Based on a Sinusoidal Representation*, Lincoln Laboratory, M.I.T., Lexington, MA, Tech. Rep. 693, 1985.