International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Time Scale Modification of Speech Signals Using Wavelet Packet Transform

N Chintaiah

Assistant Professor ,VMTW, Hyderabad, India

Abstract: ***This paper presents methods for independently modifying the time and pitch scale of acoustic signals, with an emphasis on speech signals. The algorithm developed here is based on Wavelet Packet Transform followed by sinusoidal modeling of various sub - bands. A wavelet packet tree matching to the critical bands of human ear is designed and implemented. Sinusoidal model for the speech waveform is used to develop an analysis/synthesis technique that is characterized by the amplitudes, frequencies, and phases of the component sine waves. These parameters are estimated after decomposition of the speech signal into sixteen sub-bands using wavelet packet transform. The performance of the proposed algorithm is compared with the latest method available in the literature [1] which uses Short Time Fourier Transform and sinusoidal modeling. Simulation results show that proposed method gives improved quality in time and pitch scaling compared to the other method. Performance of the proposed algorithm is demonstrated using spectrum plots also.***

Keywords: Speech analysis, speech processing, time scale modification, wavelet packet transform.

1. Introduction

There are a number of applications where it is desirable to change the time or pitch scale of an audio signal. A common instance is one in which speech needs to be slowed down in order to make it intelligible; for example 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. In both the afore-mentioned cases of rate change, it is essential that the pitch and tonal quality of the speaker should remain the same, but in others (recovery of helium-distorted speech) the pitch must be modified while the rate of articulation remains the same. Some of these methods are based on time domain splicing/overlap-add approaches which have the advantage of being computationally cheap, but at the expense of suffering from echo's.

Since that work, many other methods using the same (and related) frequency domain approaches have been developed. The algorithms involved tend to be computationally intensive, but are capable of providing very high quality output. However, they still suffer from some distortion, mainly due to the effects of „phase dispersion". That is, while the scaled signal has the same frequency content, the phases between the components change, resulting in a different wave shape.

The contribution of this paper is to develop new frequency domain type time scale modification methods that provide improved quality output by addressing the phase dispersion problem, while at the same time introducing important modifications that significantly reduce computational burdens.

Sinusoidal model using wavelet packet transform [2] for the analysis and synthesis of speech signals, estimates the sinusoidal parameters more accurately since the basis functions have compact support. In this, the authors employed a uniform wavelet packet tree structure. But, the method proposed in this paper for time and pitch scale modification, employs a wavelet packet tree structure which is non uniform and closely matching to the critical bands of the human auditory system. This filter-bank will provide better time resolution for high frequencies and better frequency resolution for low frequencies. Hence, wavelet packet transform followed by sinusoidal modeling gives better results because of its high resolution property compared to conventional model [1] in the case of time and pitch scaling of signals.

The remainder of this paper is organized as follows. In section 2, details of the sinusoidal model are described. Section 3, describes time and pitch scale transformation. In section 4, phase invariant method to avoid phase dispersion is detailed. In section 5, proposed method using wavelet packet analysis followed by sinusoidal modeling is illustrated. In section 6, experimental results & discussion are presented and in section 7 concluding remarks are provided

2. Sinusoidal Model Estimation

In the sinusoidal model [1], speech signal is represented as the sum of a finite number of sinusoids of various frequencies and is given by

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

where N is the number of sinusoids, A_k, W_k *and* Φk are the time varying amplitude, frequency and phase. Discrete Fourier transform (DFT) of overlapping frames of speech is taken first. i.e., STFT of the speech signal is taken first. Frequencies of sine wave components are estimated by the location of the peaks of the DFT magnitude function. Similarly *Ak* and Φk are estimated from the corresponding magnitude and phase of the Fourier Transform at these measured frequencies. The reconstructed signal will be

$$S_R^m(t) = \sum_{k=1}^{N(m)} A_k^m(t) \cos(\phi_k^m(t) + \Phi_k^m) \quad (2)$$

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

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

3. Time and Pitch Scaling

3.1Time Scaling

The process of time-scaling a sinusoidally modeled signal by a factor ρ involves the length TA of analysis frame and length TR of reconstruction frame to be related as
TR = ρ TA
This implies that the amplitude and frequency information at time t should be mapped to a new time t^ = ρ t , so that the scaled signal may be represented as

$$S_s^m(t^!) = \sum_{k=1}^{N(m)} A_k^m\left(\frac{t^!}{\rho}\right)\cos(\rho\phi_k^m(\frac{t^!}{\rho}) + \Phi_k^m)$$

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 < TA , where TA is the analysis frame length. Note that, as will become more evident later, the phase term $\phi_k^m(t)$ has been multiplied by ρ in order to preserve instantaneous frequency (the derivative of phase) during time scaling.

This results in „phase dispersion" (between sines) in that the reconstructed signal will contain the same frequency content as the original signal, but the relationship between the phases of the different components will have changed. During passages dominated by voiced speech, the effect of this phase dispersion is to produce an effect that may be qualitatively described as „chorusing".

A key contribution of this paper is to develop a new phase invariant method specifically designed to address this defect and hence improve the perceived quality of the time/pitch scaled signal.

3.2 Pitch Scaling

The process of pitch scaling involves every frequency track being scaled by the same constant amount σ so that the reconstructed pitch-scaled signal may be represented as

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

It also suffers from phase dispersion. So to avoid this dispersion, phase invariant method discussed below is employed.

4. Phase- Invariant Method

The motivation here is that while the sinusoidal model based methods just discussed are capable of what is considered high quality time and pitch scale transformations, they do suffer from a number of problems. In particular, during strongly voiced segments (ones containing only a few dominating sine wave components), phase dispersion producing distortion that may be described as „chorusing" is a major problem.

4.1 Pitch Estimator

Key to the development of a new method that reduces distortion due to phase dispersion is the need for an estimate of the so-called „pitch-period" on a given analysis frame. This period is defined according to an assumption that there is a dominant voicing component on a given analysis frame that produces a strong fundamental and associated harmonic elements. The pitch period is the period of this dominant fundamental component, and while other non-harmonic sinusoids will be apparent, it is reasoned that the dominant fundamental and harmonic ones will contribute most to perceptual qualities of the complete signal. In the sequel, distortion that appears as a „chorusing" effect will be reduced by synchronizing phases at certain time instants, and the latter will be defined in terms of this pitch period, so that an estimate of it is essential.

The purpose of this initial pitch estimate is to obtain initial estimates $n1^X, \ldots, nk^X$ of the harmonic spacings as follows

$$n_i^* = [\frac{f_i}{f_p^*} + \frac{1}{2}]$$

Using these quantities, the (weighted, by component magnitude) cumulative squared error in the choice of a pitch fp on the current analysis frame may be defined as

$$e(f_p) = \sum_{i=1}^{k} M_i\left(\frac{f_i}{n_i^*} - f_p\right)^2$$

The value of f_p minimizes the error and it is given by

$$f_p = \frac{1}{M}\sum_{i=1}^{k}\frac{M_i F_i}{n_i^*},$$

Pitch estimate on analysis frame is given by

$$T_P = \frac{1}{f_p}$$

4.2 Phase-Invariant Method

Phase invariant method [1] matches the set of phases in reconstruction period to phases in original signal so that phase dispersion is minimized. This method generates interpolant on reconstruction frame to maintain consistent with measured frequency information. $\phi_k^m(t)$ will be formed such that end of frame phase matching should occur. In order to reduce the phase dispersion, phases should match time points which are multiples of the pitch period T^p as shown in fig 2.

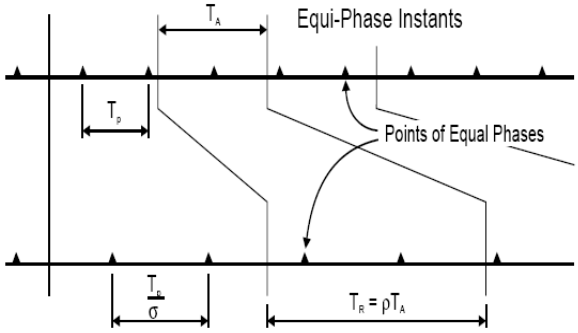

Figure 1: Example of a set of equi-phase instants

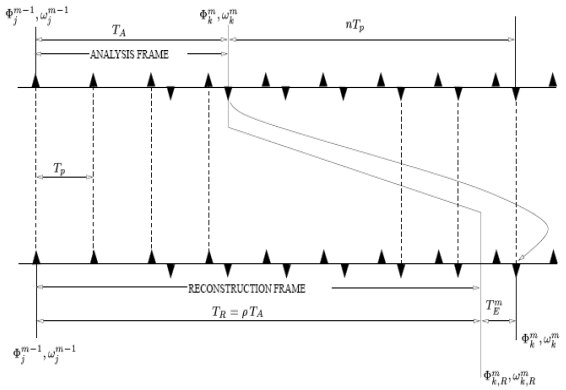

Figure 2: Interpolation criteria for generation of equi-phase $\phi_k^m(t)$

The black triangles on analysis and reconstruction frames represent the same set of phases called equi-phase. The equi phase sets are generated by adding integer multiples of $T`p$ to the original and $\frac{T_P}{\sigma}$ on the reconstruction frame. Even though starting of the analysis and reconstruction frames are in equi-phase, there may not be equi-phase at the end of the frames. But the end of the frames should be in equi-phase. So, interpolant should be chosen such that, to match the end-frame phase $\phi_k^m(t)$ at the point on the reconstruction frame which is closest to a point which is equi-phase with the end of the analysis frame as shown in fig 3. So, time and pitch scale modified signals are reconstructed by using

$$S_R^m(t) = \sum_{k=1}^{N(m)} A_K^m(t)\cos\phi_k^m(t), 0 \le t \le T_R$$

where A_k^m is the linear interpolant, ϕ_k^m is the cubic interpolant and the remaining interpolant constraints are developed .

5. Wavelet Packet Transform

In the sinusoidal model [1] described in section 2, frame size is of fixed length and is about two or more than that of an average pitch period**.** This model is not effective in achieving an optimal spectral resolution. Another problem of this is the difficulty in modeling noise-like components and time-localized transient events. These result in pre-echo distortion in the synthetic signal. Authors [2] suggested a model in which wavelet packet transform is employed instead of DFT. Figure 3 shows the frequency separation achieved by wavelet packet decomposition. Here, they used a uniform wavelet packet tree.

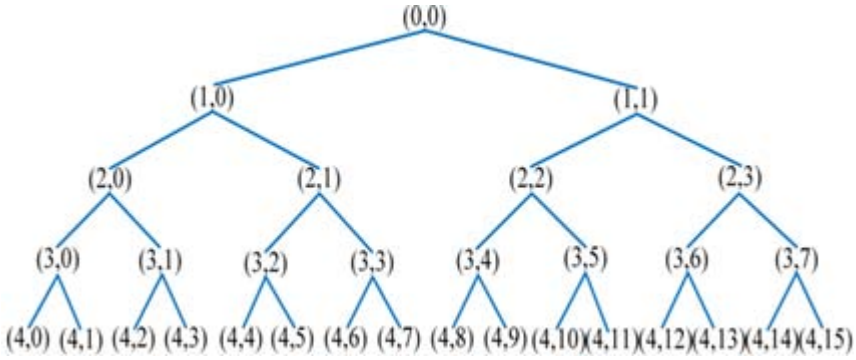

Figure 3: Wavelet packet tree and its sub-band decomposition

Above method of sinusoidal model using wavelet packet analysis [2] estimates the sinusoidal parameters more accurately since the basic functions have compact support. Instead of using a uniform wavelet packet tree, we designed a non uniform tree structure closely mimicking the critical bands of the human ear. Hence, in the proposed method, the input speech signal is first decomposed into 16 sub-bands (see figure 6) matching to the critical bands of human ear using wavelet packet transform with db4 wavelet. Even though 25 critical bands are there, only seventeen critical bands (see table 1) are sufficient in this case, since the frequency range of human speech is from 300 Hz to 3.3 kHz. It may be noted that sub-band 4 covers critical bands 4 & 5. Lower frequency parameters are calculated over a greater length of time and have higher frequency resolution. On the other hand, higher frequency sinusoidal parameters are estimated with high time resolution but poor frequency resolution.

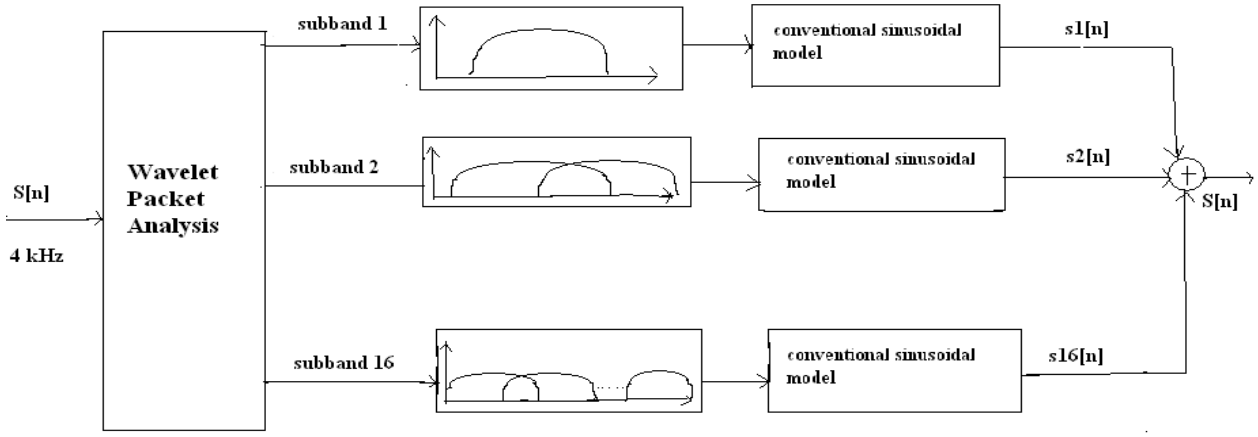

Figure 4: Block diagram of the proposed sinusoidal model using wavelet packet analysis

Table 1: First seventeen critical bands of human auditory system

Bands	*Centre Frequency*	*Band width(Hz)*
1	50	0-100
2	150	100-200
3	250	200-300
4	350	300-400
5	450	400-510
6	570	510-630
7	700	630-770
8	840	770-920
9	1000	920-1080
10	1175	1080-1270
11	1370	1270-1480
12	1600	1480-1720
13	1850	1720-2000
14	2150	2000-2320
15	2500	2320-2700
16	2900	2700-3150
17	3400	3150-3700

The block diagram of the proposed model is shown in figure 4. By using wavelet packet tree the input speech signal is decomposed into sub-band signals to analyze and synthesize each sub-band independently . The input signal is band limited to 4 kHz and sampled at 8 kHz. A Wavelet packet filter-bank is designed and implemented such that 16 sub-bands obtained closely match with the critical bands of the human auditory system.

As shown in figure 4, sinusoidal modeling of 16 sub-bands are done to obtain s1[n], s2[n], …s16[n].Amplitude, frequency and phase parameters are then estimated. These parameters are then modified to achieve time scaling and pitch scaling.

As already discussed, both STFT followed by sinusoidal modeling [1] and the proposed method suffer from phase dispersion. So, to avoid phase dispersion phase invariant method [1] is used . The performance of the proposed method is compared with the other method. It was found that, perceptual quality of the modified speech is better in the case of proposed (WPT) method. Mean Opinion Score (MOS) is calculated using the scale shown below and the results are shown in the Table2.

Mean Opinion Score

MOS	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 2: MOS obtained using [1] and the proposed method.

Model	*Time scaling of male speech*	*Pitch scaling of male speech*	*Time scaling of female speech*	*Pitch scaling of female speech*
Ref [1]	4.23	3.90	4.11	4.17
Proposed	4.87	4.61	4.77	4.73

6. Experimental Results

In this section, results of time scaling and pitch scaling on 3 seconds duration, 8 kHz sampled speech signals of male and female are shown.

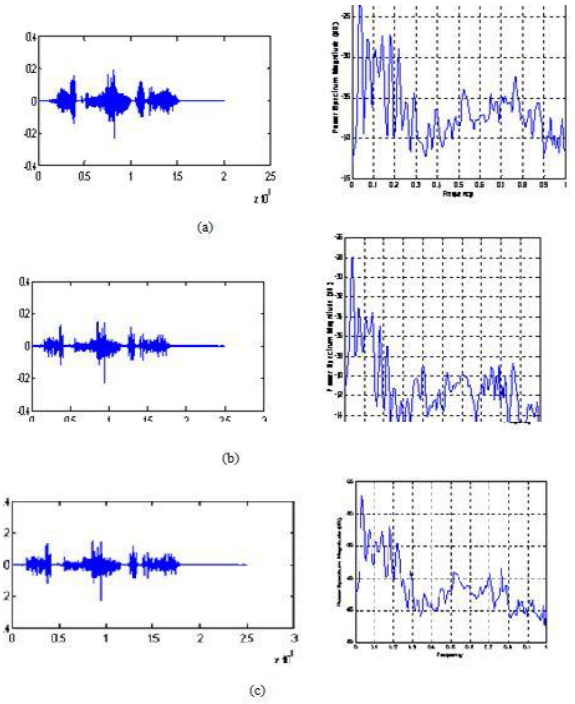

Figure 5: Time scale expansion of speech (male).(a) original with corresponding psd (b) expansion with Ref{1] (c) expansion with proposed method (ρ=1.3)

Fig 5(a) shows original male voice, 5(b) shows time scaled version of that voice with ρ= 1.3 by Ref [1] with a window size of 5ms and 5(c) shows time scaled version of the same speech using the proposed method. In figure 5, right side plots show corresponding psds. It is very clear from the power spectrum plots that there are no significant changes in the frequency contents of the signal after stretching. Hence pitch is not altered by time scaling.

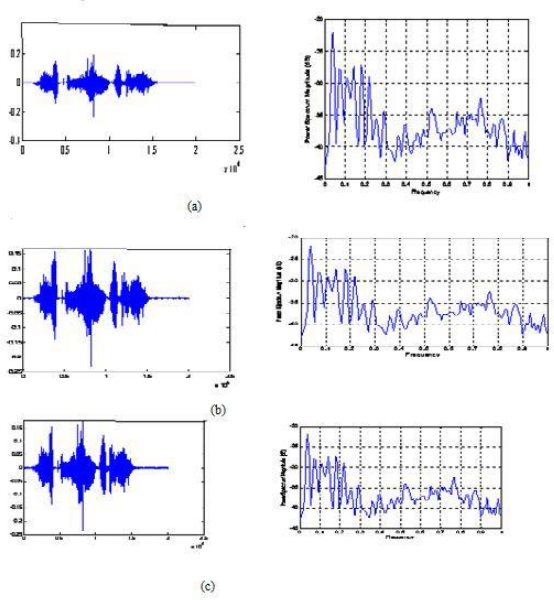

Figure 6: Pitch modification of speech(male) (a) original (b)Pitch scaled speech with Ref[1] (c) with proposed method (σ =2)

Figure 6 shows results of pitch scaling of male voice with σ = 2, and the corresponding psds. Here, it can be seen that frequency contents are changed which is an indication of pitch modification.

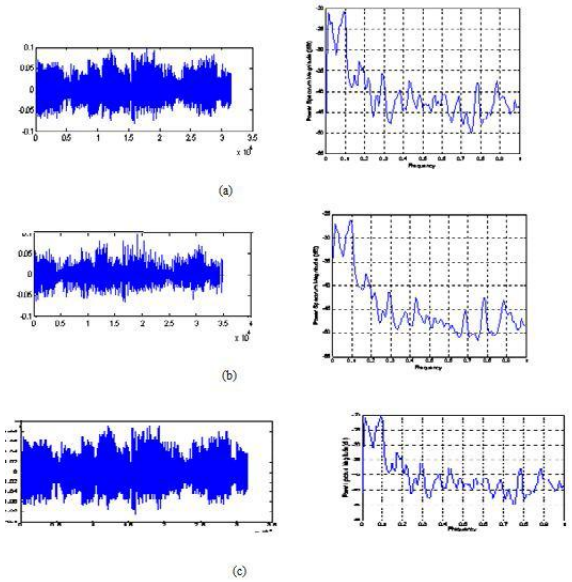

Figure 7: Time scale expansion of speech (female).(a) original with corresponding psd (b) expansion with Ref[1] (c) expansion with proposed method. (ρ = 0.7)

Figure 7 shows results of time scaling of female speech with ρ = 0.7 and Figure 9 shows the results of pitch scaling of female speech with σ = 1.5.

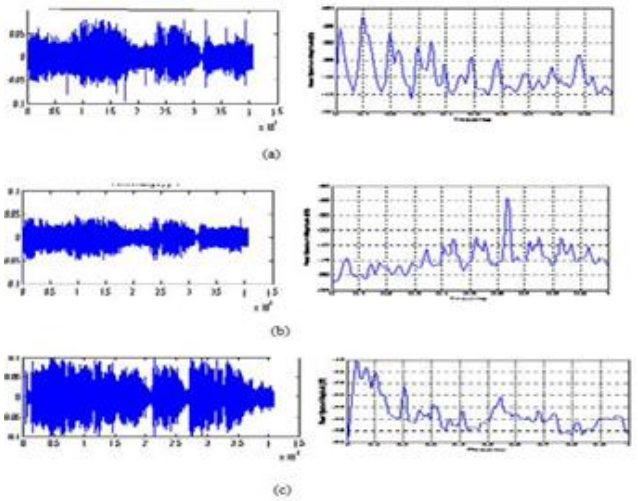

Figure 8: Pitch modification of speech(female) (a) original (b)Pitch scaled speech with Ref [1] (c) with proposed method (σ =1.5)

From these experimental results, it can be concluded that the proposed system is capable of independently modifying time scale and pitch scale of speech signals. MOS results show that the perceptual quality of modified speech signals is better than that of with the method available in the literature. This is because of the employment of wavelet packet tree matching to the critical bands of the human auditory system**.**

7. Conclusion

Methods for independently varying time scale and pitch scale of speech signals are proposed in this paper. These methods employ sinusoidal modeling of sub-bands of speech which are obtained by wavelet packet decomposition. A wavelet packet tree structure closely mimicking the critical bands of human auditory system is designed and implemented here. The performance of the proposed time scale & pitch scale methods are compared with that of methods using STFT followed by sinusoidal modeling. The scaling method uses parametric modeling techniques to achieve independent time and pitch scaling of audio signals. Phase dispersions are eliminated by employing a phase invariant method. This method does not require the decomposition of the signal into excitation and vocal tract responses. STFT followed by sinusoidal method is not effective in terms of achieving an optimal spectral resolution to each sinusoidal parameter where as wavelet packet transform followed by sinusoidal modeling gives better results because of its high resolution property.

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] Kihong Kim, Jinkeun Hong, and Jongin Lim , "Sinusoidal Modeling Using Wavelet Packet Transform Applied to the Analysis and Synthesis of Speech

Signals" V. Matouˇsek et al. (Eds.): TSD 2005, LNAI 3658, pp. 241–248, 2005.
[3] R. McAulay and T. Quatieri, "Speech analysis-syntheses based on a sinusoidal representation," IEEE Trans. Acoust., Speech., SignalProcess., vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.
[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 pitchscale 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 xpansion,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.