ABSTRACT

The paper investigates a new speech enhancement scheme to meet the demand for quality noise reduction algorithms capable of operating at a very low signal-to-noise ratio. The generalized perceptual wavelet denoising method is employed to reduce residual noise and improve the quality of speech. The proposed method is a generalized time-frequency subtraction algorithm, which advantageously exploits the wavelet multirate signal representation to preserve the critical transient information. The wavelet coefficients are used to calculate the Bark spreading energy and temporal spreading energy, from which a time-frequency masking threshold is deduced to adaptively adjust the subtraction parameters.

Keywords: Wavelet, perceptual, denoising, speech enhancement.

1. INTRODUCTION

People use speech to communicate message when speaker and listener are near to each other in a quite environment, communication is generally easy and accurate. However, in a noisy background, the listener’s ability to understand suffers. A new speech enhancement scheme for a single microphone system to meet the demand for quality noise reduction algorithms capable of operating at a very low signal-to-noise ratio. A psychoacoustic model is incorporated into the generalized perceptual wavelet denoising method to reduce the residual noise and improve the intelligibility of speech. The presented method, which advantageously exploits the wavelet multirate signal representation to preserve the critical transient information, simultaneously models the frequency and temporal masking of the human auditory system by the perceptual wavelet packet transform via the Bark and temporal localization of speech components. The wavelet coefficients are used to calculate the Bark spreading energy and temporal spreading energy, from which a time-frequency masking threshold is deduced to adaptively adjust the subtraction parameters of the proposed method. An unvoiced speech enhancement algorithm is also integrated into the system to improve the intelligibility of speech. Through rigorous objective and subjective evaluations, it is shown that the proposed speech enhancement system is capable of reducing noise with little speech degradation in adverse noise environments and the overall performance is superior to several competitive methods.

2. DWT AND FILTER BANKS

The discrete wavelet transform (DWT) corresponding to a continuous wavelet transform (CWT) function $W(a, b)$ can be obtained by sampling the co-ordinates $(a, b)$ on a grid as shown in Fig. 1. This process is called the dyadic sampling because the consecutive values of discrete scale as well as the corresponding sampling intervals differ by a factor of two. Then the dilation takes the values of the form $a = 2^k$ and translation takes the values of the form $b = 2^k l$ where $k$ and $l$ are integers. The values of $d(k, l)$ represent the discretized values of CWT $W(a, b)$ at $a = 2^k$ and $b = 2^k l$. The two-dimensional square $d(k, l)$ is commonly referred to as the discrete wavelet transform of $f(t)$. The $f(t)$ can be found using the following equation.

$$f(t) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} d(k, l) 2^{-k} \psi(2^{-l} t - l) \quad (1)$$

A. Multi-Resolution Analysis Using Filter Banks

Filters are one of the most widely used signal processing functions. Wavelets can be realized by iteration of filters with rescaling. The resolution of the signal, which is a measure of the amount of detail information in the signal, is determined by the filtering operations, and the scale is determined by up sampling and down sampling (sub sampling) operations.

The DWT is computed by successive low pass and high pass filtering of the discrete time-domain signal as shown in Fig. 1. This is called the Mallat algorithm or Mallat-tree decomposition. Its significance is in the manner it connects the continuous-time multiresolution to discrete-time filters. In the figure, the signal is denoted by the sequence $x[n]$, where $n$ is an integer. The low pass filter is denoted by $G_0$ while the high pass filter is denoted by $H_0$. At each level, the high pass filter produces detail information, $d[n]$, while the low pass filter associated with the scaling function produces coarse approximations, $a[n]$. 
At each decomposition level, the half band filters produce signals spanning only half the frequency band. This doubles the frequency resolution as the uncertainty in frequency is reduced by half. In accordance with Nyquist’s rule if the original signal has a highest frequency of $\omega$, which requires a sampling frequency of $2\omega$ radians, then it now has a highest frequency of $\omega/2$ radians. It can now be sampled at a frequency of $\omega$ radians thus discarding half the samples with no loss of information. This decimation by 2 halves the time resolution as the entire signal is now represented by only half the number of samples. Thus, while the half band low pass filtering removes half of the frequencies and thus halves the resolution, the decimation by 2 doubles the scale.

With this approach, the time resolution becomes arbitrarily good at high frequencies, while the frequency resolution becomes arbitrarily good at low frequencies. The time-frequency plane is thus resolved. The filtering and decimation process is continued until the desired level is reached. The maximum number of levels depends on the length of the signal. The DWT of the original signal is then obtained by concatenating all the coefficients, $a[n]$ and $d[n]$, starting from the last level of decomposition. Fig. 2 shows the reconstruction of the original signal from the wavelet coefficients. Basically, the reconstruction is the reverse process of decomposition. The approximation and detail coefficients at every level are up sampled by two, passed through the low pass and high pass synthesis filters and then added. This process is continued through the same number of levels as in the decomposition process to obtain and the original signal. The Mallat algorithm works equally well if the analysis filters, $G_0$ and $H_0$, are exchanged with the synthesis filters, $G_1$, $H_1$.

**A. Perceptual Wavelet Filter Bank**

To design effective algorithm for enhancing speech, it needs a well build psychoacoustic model of the ear, which has an unsurpassed capability to adapt to noise. In this section, we propose a new human auditory model that adopts the base structure of traditional auditory model but replace the time-invariant band pass filters with WPT in order to mimic the time-frequency analysis of the critical bands according to the hearing characteristics of human cochlea.

A PWPT is used to decompose the speech signal from 20 Hz to 16 kHz into 24 frequency sub bands that approximate the critical bands. The decomposition is implemented with an efficient seven-level tree structure. The analysis filter bank is depicted in Fig. 3. In the proposed wavelet packet decomposition, the two-channel wavelet filter banks are used to split both the low-pass and high-pass bands as opposed to only the decomposition of low-frequency bands in the usual wavelet decomposition. This binary tree decomposition is attractive due to the following reasons. First, the smoothness property of wavelet is determined by the number of vanishing moments. If a wavelet with a large number of vanishing moments is used, the stringent bandwidth and stop band attenuation of each sub band, as specified by the human auditory model, can be more closely approximated by the wavelet decomposition. Second, the psychoacoustic study of human ears suggests that a frequency to bark transformation needs to be performed to accurately model the frequency dependent sensitivity of human ears. Such a transformation is accomplished in audio processing systems by dividing the frequency range into critical bands.
The 24 critical bands are derived from the perfect reconstruction filter bank with finite-length filters using different wavelets for the analysis and synthesis scaling functions. Let \( H(z) \) and \( G(z) \) is the low pass (LP) and high pass (HP) transfer functions, respectively, before the decimation-by-two operation in each stage of the analysis filter bank. Further, let \( F(z) \) and \( J(z) \) is the LP and HP transfer functions, respectively, after the up sampling-by-two operation in each stage of the synthesis filter bank. The analysis and synthesis filter banks are related by

\[
g(n) = (-1)^{n} f(n) G(z) = F(-z) \tag{2}
\]

\[
j(n) = -( -1)^{n} h(n) J(z) = - H(-z) \tag{3}
\]

The relationship between the LP and HP filters reduces the number of filters to be implemented for each stage of the two-channel filter bank by half. Once the LP filters, \( H(z) \) and \( F(z) \) are designed, the HP filters, \( G(z) \) and \( J(z) \) can be derived from (2).

There are 24 critical bands in the hearing range of 0-16 kHz. Owing to the property of frequency selectivity related to critical band, temporal resolution of the human ear, and regularity property of wavelets, Daubechies wavelet basis is chosen as prototype filter and a seven-stage WPT (the number of stages \( J_{\text{max}} = 7 \), the frame length, \( F_{\text{max}} = 128 \), the frame length at stage \( j \) is given by \( F_{j} = 2^{j} \)) is adopted to build perceptual wavelet filter bank. \( W_{jk} \) represents WPT coefficient, where \( k \) is the coefficient number; \( j \) is the transform stage from which \( W_{jk} \) is chosen.

3. ADAPTIVE SPEECH ENHANCEMENT SYSTEM

The block diagram of the proposed speech enhancement system is shown in Fig. 4. After the noisy signal \( x[n] \) is decomposed by PWPT; the transform sequence is enhanced by a subtractive-type algorithm to produce the rough speech estimate. This estimate is used to calculate a time-frequency masking threshold. Using this masking threshold, a new subtraction rule, which is masking dependent, is designed to compute an estimation of the original speech. This approach assumes that the high-energy frames of speech will partially mask the input noise (high masking threshold), hence reducing the need for a strong enhancement mechanism. On the other hand, frames containing less speech (low masking threshold) will undergo an overestimated subtraction. To further improve the intelligibility of processed speech, an USE is applied. Finally, the processed speech is reconstructed by the inverse PWPT. The noise estimation is assumed to be available and is performed during speech pauses.

4. RESULTS DISCUSSION

Fig. 5 shows original speech signal which was audio file used for the processing and further simulation and experimentation. It was noiseless, clean speech signal. The noise added in the original speech signal through random signals by the White Gaussian noise to get noisy signal as shown in Fig. 6, the signal to noise ratio for noisy signal 8 dB. This signal was given as input to the modified spectral subtraction method / approach. The FFT compute the magnitude and phase spectrum of speech signal, then it was applied to magnitude spectrum of speech signal to noise estimation block for removing of noise in which by adjusting the values of \( \beta \), min. and max. SNR and also min. and max. Alpha. The Fig. 7 and Fig. 8 shows the spectrogram of noisy signal and original signal respectively which could be used to understand the distribution of energy. The Fig. 9 shows the decomposition of signal using haar and Daubechies wavelet transform. From the Fig. 4 the the o/p of noise spectrum estimation was subtracted to get the noise parameters from the magnitude spectrum of noisy signal. For the improvement and time-frequency filtering purpose it was applied to spectrally subtracted noise free speech signal to additional noise suppression and smoothing block respectively as shown in Fig. 4 of modified spectral subtraction approach. For getting sequence of speech signal output of smoothing was applied to overlap and add block which provide enhanced speech signal.

The signal is given as input to the inverse perceptual wavelet filter bank (IPWPT) was synthesized and reconstructed using the wavelet filter bank as Inverse Perceptual wavelet transform is used to regenerate the Enhanced Speech Signal. The Enhanced Speech Signal is as shown in Fig. 17.
Fig. 7: Spectrogram of Noisy Signal

Fig. 8: Spectrogram of Original Signal

Fig. 9: Spectrogram of Enhanced Signal

Fig. 10: Decomposed Signal at D12, D6, and D36

Fig. 11: Decomposed Signal at A21, D22, and A11

Fig. 12: Decomposed Signal at D38, A19, and D20

Fig. 13: Decomposed Signal at D38, A89, and D70

Fig. 14: Decomposed Signal at A133, A134, and
5. CONCLUSION

We present the wavelet based speech enhancement method which is generalized perceptual wavelet denoising method to reduce the residual noise and improve the quality of speech. The results shows the speech is enhance at significant level. The uniform energy distribution of signal is also observed.

REFERENCES


