Speech Compression for Better Audibility using Wavelet Transformation with Adaptive Kalman Filtering

Siva Nagu.T
Department of Electronics & Communication Engineering
GIET Affiliated to JNTUK
Rajahmundry

K.Jyothi
Associate professor
Department of Electronics & Communication Engineering
GIET Affiliated to JNTUK
Rajahmundry

V.Sailaja, Ph.D
Professor
Department of Electronics & Communication Engineering
GIET Affiliated to JNTUK
Rajahmundry

ABSTRACT
In mobile communication systems, service providers are continuously met with the challenge of accommodating more users within a limited allocated bandwidth. For this reason, manufactures and service providers are continuously in search of low bit-rate speech coders that deliver toll-quality speech**.**
This paper deals with speech compression based on discrete wavelet transforms and Adaptive Kalman filter. We used English words for this experiment. We could successfully compressed and reconstructed the words with perfect audibility by using above technique. Speech compression is the technology of converting human speech into an efficiently encoded representation that can later be decoded to produce a close approximation of the original signal. The wavelet transform of a signal decomposes the original signal into wavelets coefficients at different scales and positions. These coefficients represent the signal in the wavelet domain and all data operations can be performed using the corresponding wavelet coefficients.
In this paper Code was simulated using MATLAB. The result obtained from Wavelet Coding was compared with Adaptive Kalman with Wavelet Coding. From the results we noticed that the performance of Wavelet Coding with Adaptive Kalman Filter is better than wavelet transform.

Keywords
Wavelet Transform coding (DWT), Adaptive Kalman filtering.

1. INTRODUCTION
Speech is a very basic way for humans to convey information to one another. With a bandwidth of only 4 kHz, speech can convey information with the emotion of a human voice. People want to be able to hear someone's voice from anywhere in the world, as if the person was in the same room. As a result a greater emphasis is being placed on the design of new and efficient speech coders for voice communication and transmission; today applications of speech coding and compression have become very numerous. Many applications involve the real time coding of speech signals, for use in mobile satellite communications, cellular telephony, and audio for videophones or video teleconferencing systems. Other applications include the storage of speech for speech synthesis and playback, or for the transmission of voice at a later time. Some examples include voice mail systems, voice memo wristwatches, voice logging recorders and interactive PC software.
Traditionally speech coders can be classified into two categories: waveform coders and analysis/synthesis vocoders (from .voice coders.). Waveform coders attempt to copy the actual shape of the signal produced by the microphone and its associated analogue circuits [1]. A popular waveform coding technique is pulse code modulation (PCM), which is used in telephony today. Vocoders use an entirely different approach to speech coding, known as parameter coding, or analysis/synthesis coding where no attempt is made at reproducing the exact speech waveform at the receiver, only a signal perceptually equivalent to it. These systems provide much lower data rates by using a functional model of the human speaking mechanism at the receiver. One of the most popular techniques for analysis/synthesis coding of speech is called Linear Predictive Coding (LPC).
Some higher quality vocoders include RELP (Residual Excited Linear Prediction) and CELP (Code Excited Linear Prediction) [2].
Very simply wavelets are mathematical functions of finite duration with an average value of zero that are useful in representing data or other functions. Any signal can be represented by a set of scaled and translated versions of a basic function called the mother wavelet. This set of wavelet functions forms the wavelet coefficients at different scales and positions and results from taking the wavelet transform of the original signal. The coefficients represent the signal in the wavelet domain and all data operations can be performed using just the corresponding wavelet coefficients [3].
Whispered speech is playing a more and more important role in the widespread use of mobile phones for private communication than ever. Speaking loudly to a mobile phone in public places is considered a nuisance to others and conversations are often overheard. Since noisy signals are not available directly here we are taking the original signal and adding noisy signals such as babble, car, street. Different methods such as Weiner, MMSE, spectral subtraction, wavelets are used to filter the signals from noise. These methods are used earlier but output after filtering is not accurate. So in this paper we proposed Kalman filter method which improves signal to noise ratio (SNR) of original speech compared to above methods.
This paper is organized as follows: Section 2 covers discrete wavelet transform. Section 3 covers Speech Enhancement and Kalman filtering method, Section 4 discuses Performance measurements of wavelets, Section 5 shows results. Finally, section 6 gives Conclusion.

2. SPEECH COMPRESSION USING DISCRETE WAVE TRANSFORM
Speech compression using discrete wave transforms (DWT) is shown in steps below.

2.1 Choice of Appropriate Wavelet

The choice of the mother wavelet plays a very important role in designing high quality speech. Choosing the appropriate wavelet will maximize the SNR and minimizes the relative error. Here we selected db20 wavelet for better results.

Wavelets with more vanishing moments provide better reconstruction quality, as they introduce less distortion into the processed speech and concentrate more signal energy in a few neighboring coefficients. However the computational complexity of the DWT increases with the number of vanishing moments and hence for real time applications it is not practical to use wavelets with an arbitrarily high number of vanishing moments [4].

2.2 Decomposition Level

Wavelets work by decomposing a signal into different frequency bands and this task is carried out by choosing the wavelet function and computing the discrete wavelet transform (DWT) [5]. Choosing a decomposition level for the DWT Usually depends on the type of signal being analyzed.

2.3 Truncation of Coefficients

The coefficients obtained after applying DWT on the frame concentrate energy in few neighbors. Here we are truncating all coefficients with "low" energy and retain few coefficients holding the high energy value. Two different approaches are available for calculating thresholds.

2.4 Global Thresholding

The aim of Global thresholding is to retain the largest absolute value coefficients. In this case we can manually set a global threshold. The coefficient values below this value should be set to zero, to achieve compression.

2.5 Level dependent Thresholding

This approach consists of applying visually determined level dependent thresholds to each de composition level in the Wavelet Transform. The value of the threshold applied depends on the compression. The task is to obtain high compression and an acceptable SNR needed to reconstruct the signal and detect it. Among these two, high SNR is achieved using global thresholding compared to level dependent thresholding.

2.6 Encoding

Signal compression is achieved by first truncating small valued coefficients and then efficiently encoding them.

One way of representing the high-magnitude coefficients is to store the coefficients along with their respective positions in the wavelet transform vector [5]. For a speech signal of frame size F, taking the DWT generates a frame of size T, slightly larger than F. If only the largest L coefficients are retained, then the compression ratio C is given by: $C = F/2L$

Another approach to compression is to encode consecutive zero valued coefficients [6], with two bytes. One byte to indicate a sequence of zeros in the wavelet transforms vector and the second byte representing the number of consecutive zeros.

3. SPEECH ENHANCEMENT

3.1 Modeling noisy speech and filtering

If the clean speech is represented as x(n) and the noise signal as v(n), then the noise-corrupted speech y(n), which is the only observable signal in practice, is expressed as

$$Y(n) = x(n) + v(n) \qquad (1)$$

In Wiener filtering method filtering depends on the adaptation of the transfer function from sample to sample based on the speech signal statistics (mean and variance). It is implemented in time domain to accommodate for the varying nature of the speech signal. The basic principle of the Wiener filter is to obtain an estimate of the clean signal from that corrupted by additive noise. This estimate is obtained by minimizing the Mean Square Error (MSE) between the desired signal $s(n)$ and the estimated signal $\hat{s}(n)$. Transfer Function in frequency domain is given below

$$H(\omega) = P_s(\omega) + P_v(\omega)$$

Where $P_s(\omega)$ and $P_v(\omega)$ are the power spectral densities of the clean and the noise signals, respectively.

An improved method is based on minimum mean square error-short time spectral amplitude (MMSE-STSA) is proposed to cancel background noise in whispered speech. Using the acoustic character of whispered speech, the algorithm can track the change of non-stationary background noise effectively. Compared with original MMSE-STSA algorithm and method in selectable mode Vo-coder (SMV), the improved algorithm can further suppress the residual noise for low signal-to-noise radio (SNR) and avoid the excessive suppression. Whereas in Spectral subtraction based speech enhancement methods are known to be effective for the suppression of additive stationary, broadband noise. Tonal noises such as car horn sounds are found to cause serious degradation of the output speech quality. And in wavelet de-noising method is a nonlinear de-noising method based on the wavelet decomposition. Compared with the traditional low pass filters, the wavelet de-noising method can not only realize the function of low pass filter but also maintain the feature of the signal. Among the different methods of wavelet de-noising, the wavelet threshold de-noising method is applied widely and can meet the needs of real time.

3.2 Kalman Filtering Method

The Kalman filter is an unbiased, time-domain, linear minimum mean squared error (MMSE) estimator, where the enhanced speech is recursively estimated on a sample-by-sample basis. Hence, the Kalman filter can be viewed as a joint estimator for both the magnitude and phase spectrum of speech, under non-stationary assumptions [7]. This is in contrast to the short-time Fourier transform (STFT)-based enhancement methods, such as spectral subtraction, Wiener filtering, and MMSE estimation [8], where the noisy phase spectrum is combined with the estimated clean magnitude spectrum to produce the enhanced speech frame. However, it has been reported that for spectral SNRs greater than approximately 8 dB, the use of unprocessed noisy phase spectrum does not lead to perceptible distortion [8]. Kalman filter is also used by Stephen So, Kamil K. W´ojcicki, Kuldip K. Paliwal for speech enhancement in their paper "Single-channel speech enhancement using Kalman filtering in the modulation domain" 2010 [9].Since kalman filter is a joint magnitude and phase speech estimator, which is highly suited for modulation domain processing, as modulation phase tends to contain more speech information than acoustic.

Using scalar Kalman filter for speech enhancement, v (n) is a zero-mean, white Gaussian noise that is uncorrelated with x(n). A p^{th} order linear predictor is used to model the speech signal:

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + w(n) \qquad (2)$$

Where $\{a_k, k = 1, 2. . . p\}$ are the wavelets and w(n) is the white Gaussian excitation with zero mean and a variance of σ_w .Rewriting Eq. (1) and (2) using state vector representation:

$$x(n) = Ax(n-1) + dw(n) \qquad (3)$$

$$y(n) = c^T x(n) + v(n) \qquad (4)$$

where $\mathbf{x}(n) = [x(n), x(n-1), . . . , x(n-p+1)]^T$ is the 'hidden 'state vector, $\mathbf{d} = [1, 0, . . . , 0]^T$ and $\mathbf{c} = [1, 0, . . . , 0]^T$ are the measurement vectors for the excitation noise and observation, respectively. The linear prediction state transition matrix **A** is given by:

$$A = \begin{bmatrix} -a_1 & -a_2 & \dots & -a_{p-1} & -a_p \\ 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix} \qquad (5)$$

When provided with the current sample of corrupted speech y(n), the Kalman filter calculates $\hat{\mathbf{x}}(n|n)$, which is an unbiased, linear MMSE estimate of the state vector **x**(n), by using the following recursive equations

$$P(n|n-1) = AP(n-1|n-1)A^T + \sigma_w^2 dd^T$$

$$K(n) = P(n|n-1)c[\sigma_v^2 + c^T P(n|n-1)c]^{-1}$$

$$\hat{x}(n|n-1) = Ax(n-1|n-1)$$

$$P(n|n) = [I - K(n)C^T]P(n|n-1)$$

$$\hat{x}(n|n) = \hat{x}(n|n-1) + K(n)[y(n) - C^T\hat{x}(n|n-1)]$$

The current estimated sample is then given by $\hat{x}(n) = c^T\hat{x}(n|n)$

This extracts the first component of the estimated state vector. During the operation of the Kalman filter, the noise corrupted speech y(n) is windowed into non-overlapped, short (e.g. 20 ms) frames, on them wavelets were applied and excitation variance σ2 is estimated.

4. WAVELETS PERFORMANCE MEASURES

The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently. The DWT coefficient d_{jk} are defined by Eq (6).Where Ψ (t) is the mother wavelet.

$$d_{jk} = \int X(t)\, \emptyset_{jk}(t)\, dt = 2^{j/2} \int x(t)\, \emptyset_{jk}(2^j t-k)\, dt$$

$$\emptyset_{Jk}(t) = 2^{j/2}\, \emptyset_{jk}(2^j t-k), \quad j, k \in Z \qquad (6)$$

$$X(t) = \sum_j \sum_k d_{jk}\, \emptyset_{jk}(t)$$

Where Ψ (t) is the mother wavelet and X (t) is the time signal.

A number of quantitative parameters can be used to evaluate the performance of the wavelet based speech coder, in terms of both reconstructed signal quality after decoding and compression scores.

The following parameters are compared:

- Signal to Noise Ratio (SNR),
- Peak Signal to Noise Ratio (PSNR),
- Normalized Root Mean Square Error (NRMSE),
- Percentage of zero coefficients (PZEROS)
- Compression Score (CS)

The results obtained for the above quantities are calculated using the following formulas

4.1 Signal to Noise Ratio (SNR)

This value gives the quality of reconstructed signal. Higher the value, the better:

$$\text{SNR} = 10 \log_{10}\left(\frac{\sigma x^2}{\sigma e^2}\right)$$

σx^2 Is the mean square of the speech signal and σe^2 is the mean square difference between the original and reconstructed signals.

4.2 Peak Signal to Noise Ratio (PSNR)

$$\text{PSNR} = 10 \log_{10} NX^2 / \|X - r\|^2$$

N is the length of the reconstructed signal, X is the maximum absolute square value of the signal x and ||x-r||2 is the energy of the difference between the original and reconstructed signals.

4.3 Normalized Root Mean Square Error (NRMSE)

$$\text{NRMSE} = \text{sqrt}[(x(n) - r(n))^2 / (x(n) - \mu x(n))^2]$$

Where X(n) is the speech signal, r(n) is the reconstructed signal, and x(n) is the mean of the speech signal.

4.4 Percentage of zero coefficients (PZEROS)

It is given by the relation:

PZEROS = 100*(No of Zeros of the current decomposition) / No of coefficients.

4.5 Compression Score (CS)

It is the ratio of length of the original signal to the compressed signal.

C = Length(x(n)) / Length (cWC)

cWC is the length of the compressed wavelet transform vector.

4.5 Effects of Threshold

In this experiment, there is a need to study the effects of varying threshold value on the speech signals in terms of SNR and compression score. For db20 at level 2, the threshold value was slowly increased, and the corresponding values of the SNR and Compression score were recorded in Tables 1 and 2:

Table 1. Male

Threshold Values	SNR	Compression Score(CS)
2	4.81	45.88
4	4.83	45.29
6	4.82	45.22
8	4.89	45.16

Table 2. Female

Threshold Values	SNR	Compression Score (CS)
2	3.19	37.01
4	3.20	36.86
6	3.23	37.04
8	3.19	37.51

5. RESULT

As shown in table 1 and 2 a speech files spoken in English language is recorded for both male and female. The effects of varying threshold value on the speech signals in terms of SNR and compression score were observed at different levels.

There are many factors which affects the wavelet based speech Coder's performance, mainly what compression ratio could be achieved at suitable SNR value with low value of NRMSE. To improve the compression ratio of wavelet-based coder, we have to consider that it is highly speaker dependent and varies with his age and gender. That is low speaking speed cause high compression ratio with high value of SNR. Increasing the scale value (NRMSE) in wavelet-based speech coder gives higher compression ratios.

From table 3, it is observed that kalman filter with wavelet coding has the better peak signal to noise ratio (PSNR) than wavelet transform.

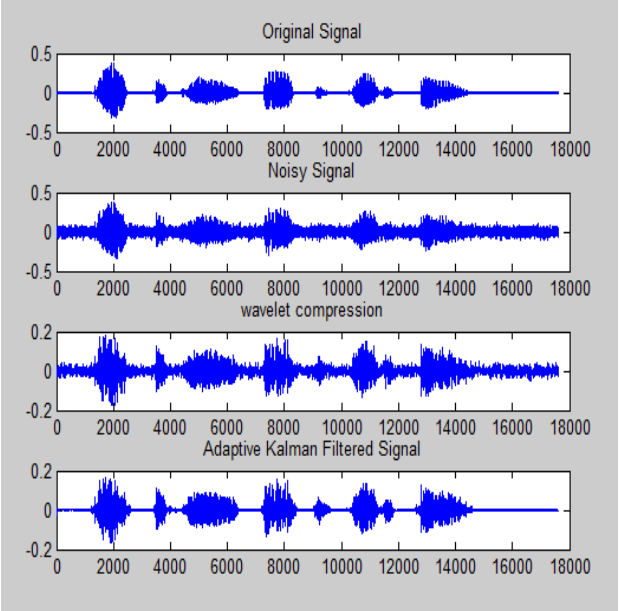

Figure 1: Output Waveform

Table 3. Noisy Speech Model with Kalman Filtering and Wavelet Coding

Wavelet	SNR	NRMSE	PZEROS	CS	PSNR WITH WAVET	PSNR WITH KALMAN
Haar	4.85	0.75	75	45.27	12.96	14.33
Sym2	6.00	0.70	74.99	50.48	13.53	14.71
Sym5	6.06	0.70	74.98	49.90	13.56	14.51
Coif2	6.06	0.70	74.98	49.49	13.56	14.51
Db20	6.13	0.70	74.94	50.45	13.60	14.46

6. CONCLUSION

Adaptive kalman filter with wavelet transformation for compressing one-dimensional signals (as speech signal) was developed. It compacts much of the signal energy into as few coefficients as possible. These coefficients are preserved and the other coefficients are discarded with little loss in signal quality.

As previously mentioned, the purpose of this approach is to reconstruct an output speech signal by making use of the accurate estimating ability of the Kalman filter.

Performance of the wavelet coder is tested on male and female speech signals of duration 10Sec. Results illustrate that the performance of Wavelet Coding with Adaptive Kalman Filter was better than wavelet transform.

7. REFERENCES

[1]. J.N. Holmes, Speech Synthesis and Recognition, Chapman & Hall, London, 1988.

[2]. A. Gersho, "Speech Coding," Digital Speech Processing, A.N. Ince, ed., Kluwer Academic Publishers, Boston, 1992, pp. 73-100.

[3]. Hatem Elaydi, Mustafa I. Jaber, Mohammed B. Tanboura, "Speech Compression using Wavelets" Electrical & Computer Engineering Department Islamic University of Gaza, Palestine,2010.

[4]. V. Viswanathan, W. Anderson, J. Rowlands, M. Ali and A. Tewfik, "Real-Time Implementation of a Wavelet-Based Audio Coder on the T1 TMS320C31 DSP Chip," 5th International Conference on Signal Processing Applications & Technology (ICSPAT), Dallas, TX, Oct. 1994.

[5]. E.B. Fgee, W.J. Phillips, W. Robertson, "Comparing Audio Compression using Wavelets with other Audio Compression Schemes," IEEE Canadian Conference on Electrical and Computer Engineering, IEEE, Edmonton, Canada, 1999, pp. 698-701.

[6]. W. Kinsner and A. Langi, "Speech and Image Signal Compression with Wavelets," IEEE Wescanex Conference Proceedings, IEEE, New York, NY, 1993, pp. 368-375.

[7]. C. J. Li, "Non-Gaussian, non-stationary, and nonlinear signal processing methods – with applications to speech processing and channel estimation," Ph.D. dissertation, Aarlborg University, Denmark, Feb. 2006.

[8]. P. Loizou, Speech Enhancement: Theory and Practice, 1st ed. CRC Press LLC, 2007.

[9]. Stephen So, Kamil K. W´ojcicki, Kuldip K. Paliwal "Single-channel speech enhancement using Kalman filtering in the modulation domain" Sept. 2010, Signal Processing Laboratory, Griffith School of Engineering, Griffith University, Brisbane, QLD, Australia, 4111