Grenze International Journal of Computer Theory and Engineering , June 2017

A Pattern Comparision Approach to Speech versus Silence Discrimination as a Preprocessing Front End for Speech Recogition

Mahadevaswamy[1] and D J Ravi[2]
[1]Assistant Professor, Vidyavardhaka College of Engineering, Mysuru,
Visvesvaraya Technological University, Belagavi,
Email: mahadevaswamy@vvce.ac.in
[2]Professor, Vidyavardhaka College of Engineering, Mysuru,
Visvesvaraya Technological University, Belagavi,
Email: ravidj@vvce.ac.in

***Abstract*—The computational performance of automatic speech recognition can be enhanced either at the front end by measuring a set of patterns for speech versus silence or noise discrimination, also by speech enhancement or by an efficient classifier at the backend. Employing the speech versus silence or noise discrimination before performing speech enhancement operations plays a vital role, to prevent the over thresholding of silence coefficients to reduce the computational complexity that would not be useful for maintaining the intelligibility of speech. The comparison of experimental results on the database created in the room environment and on the standard TIMIT database reveal that the performance of the proposed system is better under noisy as well as in noise free conditions.**

***Index Terms*— Wavelet Packet Decomposition, Visu Shrink, Teager Energy Operator, Time Adaptive Threshold.**

I. INTRODUCTION

Speech versus silence or noise discrimination is an important step in automatic speech recognition systems that aim to effectively separate speech from background noise or silence. Speech versus noise discrimination is a preprocessing front end generally followed by powerful speech enhancement techniques, robust feature extraction and automatic pattern classification in majority of the speech processing systems. The tremendous advancement of new technology and solutions has made the man machine communication tasks like automatic speech recognition system a reality. But, the most familiar fact is that these systems perform well in the clean conditions, in the presence of stationary noise and fail to perform well in the presence of nonstationary noise. Another meaningful cause for poor performance of the automatic speech recognition system is mainly due to the mismatch in the training and testing environments. So, in order to improve the performance of these systems we need to use speech versus noise discrimination, speech enhancement, efficient feature extraction and classification. The traditional methods for speech versus noise discrimination

Grenze ID: 01.GIJCTE.3.2.6

and enhancement are spectral subtraction[10], MMSE[7], subspace method[8], Wavelet Domain based[5] and Wiener filtering[3]. The most familiar method is the wavelet denoising proposed by Donoho and Johnstone. This method is based on the thresholding[9], [11] of coefficients of the noisy speech to reduce the noise content. Wavelet means short duration wave. The general principle of wavelet based enhancement techniques is either to include or discard wavelet packet coefficients using based on the desired adaptive threshold and appropriate thresholding techniques. Since voiced, unvoiced sound has large number of approximation and detail coefficients which are similar, addition of background noise or silence to the speech signal in the training and testing environments leads to instant reduction in the intelligibility of speech. To prevent this we need to use speech versus noise discrimination and enhancement. Thus the quality of speech sound is enhanced. The organization of the paper is as follows, the section II gives brief literature survey about traditional speech enhancement methods. Section III describes about the methodology followed in this paper. Results are presented in section IV and finally concluded in section V.

II. Literature Survey

Tavares et al., presented a scheme for speech enhancement by calculating the noise standard deviation from distributions of signal and noise for choosing of adaptive threshold [12]. Sanjay P. et. al. explored a speech enhancement principle using the harmonic model to calculate the frame to frame phase difference in voiced segments and unvoiced segments. This method was used with spectral subtraction and log-MMSE STSA for noise reduction in presence of the non stationery noise[1]. Sanam et. al. discussed an improved wavelet packet scheme based on speech enhancement using semisoft threshold. They summarized the future scope as experimenting with the different distributions and use of perceptual wavelets[2]. Bingyin Xia et. all. proposed a novel speech enhancement technique based on Weighted Denoising Auto-encoder(WDA). Here the clean speech is determined by Weiner filter. The technique gave better results than Weiner filtering method based on decision directed approach[3]. Sanam et al., explored a speech enhancement method using the distribution of probability and threshold estimated through symmetric K-L divergence. Performance is analyzed and evaluated using metrics SNR, PESQ, WSS through a comparative study[4]. Chen et al., Presented the denoising problem by improvising the performance of speech enhancement through PWPD, TEO in the wavelet transform domain[5]. Chang et al., proposed Bayes shrink for speech enhancement and its performance is evaluated through comparative study with Sureshrink[6]. Donoho et al., proposed universal threshold for speech enahncement[11]. J F Kaiser proposed TEO to determine the energy of complex functions[13]. Ephraim Y et al. explored short time spectral amplitude based denoising algorithm and summarized that the performance of the MMSE STSA is comparable to that of Wiener STSA at higher SNR values except that MMSE STSA has minimum mean square error at low SNR values[7]. They have also proposed a signal subspace approach for speech denoising and its performance is compared with spectral subtraction method[8]. D L Donoho proposed a wavelet based soft thresholding technique using semi soft threshold for denoising[9].

III. Proposed Method

The preprocessing steps in the robust automatic speech recognition algorithms are speech discrimination from silence or background noise, speech enhancement. The wavelet based denoising is a nonlinear speech enhancement procedure, where the main intention is to find an estimate of the input signal from degraded speech $X(n)$.

$$\mathrm{X}(n) = x(n) + N(n) \tag{1}$$

where$N(n)$ indicates a Additive white Gaussian Noise(AWGN). The SVSD and adaptive wavelet denoising includes the following steps.

- Measurement of reference and test features.
- Comparison of test features with the reference features.
- Decision for Speech Discrimination from noise or silence.
- Wavelet Packet Decomposition (WPD), Wavelet Packet Reconstruction (WPR).
- Thresholding.

- Signal estimation from noise.

The block diagram of the speech versus noise discrimination and adaptive wavelet based denoising is as shown in figure 1. The detailed description about each step is as follows.

A. Pattern Measurements

A pattern recognition approach is used to distinguish the given speech sound from the noisy background or from the silence. This is achieved by measuring the patterns such as zero crossing rate, energy and correlation coefficient. The patterns are measured for both training and testing inputs, before employing the given speech segment as input to the enhancement technique. The figure 1 describes the speech enhancement system based on the principle of speech versus noise discrimination. The details of measured patterns are as follows.

- The number of zero crossing count for a speech segment $x(n)$ is given by the equation,

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]| \qquad (2)$$

where $sgn[x(n)] = \begin{cases} -1 & x(n) < 0 \\ 1 & x(n) \geq 0 \end{cases}$ (3)

- The energy of the signal $x(n)$ is defined by

$$E_s = \left(\sum_{n=1}^{N} x^2(n)\right) \qquad (4)$$

Where N is the number of samples in the given speech utterance.

- The Normalized correlation coefficient at unit time delay is given by

$$C_1 = \left(\frac{\sum_{n=1}^{N} x(n)x(n-1)}{\sqrt{\sum_{n=1}^{N} x^2(n) \sum_{n=0}^{N-1} x^2(n)}}\right) \qquad (5)$$

B. Speech Versus Silence Discrimination(SVSD)

The pattern measurements of the training and test input speech sound made in the preceding step are used to make decision through pattern recognition to differentiate the given speech sound from the noisy background or silence class. To make decision the patterns of test input are compared against patterns of trained input. Then based on the result the test data is assigned to class i, where $i = 1$, for speech class, $i = 2$ for Background noise or silence class. The figure 2 indicates the block diagram proposed speech versus noise or silence discrimination method.

C. Wavelets[5][14]

Wavelet is of very short duration signal. The principle of wavelets is that creating a filter by performing, dilation and transform operations on the mother wavelet. Hence all wavelets are the dilated and transformed version's of standard wavelet known as mother wavelet. The following equation describes the wavelet in continuous time domain.

$$CWT(a, b, f(t)) = \int_{-\infty}^{\infty} f(t)\, \varphi^{*}{}_{a,b}\, dt \qquad (6)$$

$$\varphi_{a,b} = \frac{1}{\sqrt{a}} \varphi\left(\frac{t-b}{a}\right) \qquad (7)$$

Where a and b are real numbers, $\varphi_{a,b}$ is mother wavelet, $f(t)$ is the input signal and * indicates that it is complex conjugate operation. The compressing and expanding operation leads to compressed wavelets ($a < 1$) and dilated wavelets ($a > 1$). We need to apply compressed wavelet to high frequency segments and stretched wavelet to low frequency segments in the signal respectively. This is essential to prevent the loss of information. The discrete wavelet transform is given by

$$d_{j,k} = \int_{-\infty}^{\infty} f(t)\, \varphi_{j,k}^{*}(t) dt \qquad (8)$$

$$\varphi_{j,k}(t) = {a_0}^{-\frac{j}{2}} \varphi\left({a_0}^{-j} t - kT\right) \qquad (9)$$

selection of j, k based on dyadic scales and positions (a_0= 2), then dwt is given by

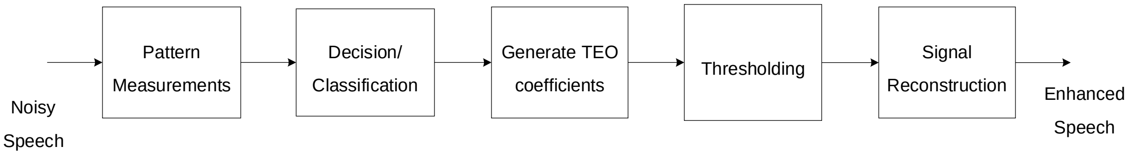

Figure 1. The block diagram of speech versus silence discrimination based speech enhancement

$$d_{j,k} = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} f(t)\, \varphi_{j,k}^{*}\left(\frac{t}{2^j} - kT\right) dt \qquad (10)$$

where j specifies the subband level and k specifies the subband number, $f(t)$ is input signal, T specifies the time period.

Analysis & Synthesis using wavelet packet decomposition algorithm

The wavelet packets were first introduced by Coifman. The noisy speech is subjected to n-Level Wavelet Packet Decomposition[5] to decompose into approximation and detail coefficients $\omega_{j,m}(k)$.The desired operations are applied to the coefficients. The n-Level Wavelet Packet Reconstruction is employed to obtain an estimate of the input signal.

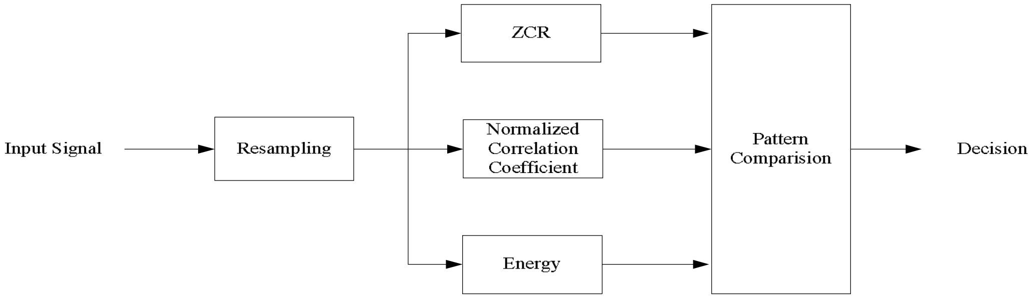

Figure 2. Block Diagram of the Analysis System

TEO coefficients[13]

The Teager Energy Operator (TEO) is proposed by Kaiser. The TEO coefficients are determined by employing teager energy operator on the approximation and detail coefficients in discrete time domain. The TEO is given by,

$$\psi[x(n)] = x^2(n) - x(n+1)x(n-1) \qquad (11)$$

Where $\psi[.]$ is the TEO operator, $x(n)$ is speech signal. TEO coefficients are given by

$$T_{j,m}(k) = \Psi\left[w_{j,m}(k)\right] \qquad (12)$$

Where $w_{j,m}(k)$are wavelet co-efficients and $T_{j,m}(k)$ are TEO coefficients.

Universal threshold

The standard deviation σ_j of noise is used to obtain the constant threshold[5][11].

$$\lambda_j = \sigma_j\sqrt{2\log N} \qquad (13)$$

N indicates the number of samples. The temporal masking is determined by using the equation [5]

$$M_{j,m}(k) = T_{j,m}(k) * H_j(k) \qquad (14)$$

where $*$ denotes the convolution operation, $T_{j,m}(k)$ are TEO coefficients and $\mathrm{H_j(k)}$ is the hamming window of

length = $256/2^j$. The modulated temporal masking construction[5] is given by

$$M'_{j,m}(k) = M_{j,m}(k) \qquad (15)$$

TAT using universal threshold is given by

$$\lambda_{j,m}(k) = \lambda_j\left(1 - M'_{j,m}(k)\right) \qquad (16)$$

Hard Thresholding Technique[5]

Hard thresholding results in

$$\widehat{\omega}_{j,m}(k) = \begin{cases} \omega_{j,m}(k) & \left|\omega_{j,m}(k)\right| \geq \lambda_{j,m}(k) \\ 0 & \left|\omega_{j,m}(k)\right| < \lambda_{j,m}(k) \end{cases} \qquad (17)$$

Where $\widehat{\omega}_{\mathrm{j,m}}(\mathrm{k})$ is the m^{th} threshold operated coefficients of k-th subband.

Soft Thresholding Technique[5][9]

The soft thresholding is described by the following equation

$$\widehat{\omega}_{j,m}(k) = \begin{cases} sgn\left(\omega_{j,m}(k)\right)\left(\left|\omega_{j,m}(k) - \lambda_{j,m}(k)\right|\right) if \left|\omega_{j,m}(k)\right| \geq \lambda_{j,m}(k) \\ \quad 0, \qquad\qquad \text{otherwise} \end{cases} \qquad (18)$$

where $\widehat{\omega}_{j,m}(k)$ is the thresholded wavelet coefficients of the m^{th} subband.

E. *Wavelet Packet Reconstruction[5]*

An approximate version of the original speech is found by using wavelet packet reconstruction.

IV. Experimental Results

The SVSD wavelet denoising performance is evaluated using the signal to noise ratio[5] i.e.,

$$SNR = 10\log_{10}\left(\frac{\sum_{n=1}^{N}\{|s(n)|^2\}}{\sum_{n=1}^{N}\{|s(n) - \hat{s}(n)|^2\}}\right)[dB] \qquad (19)$$

A. *Database*

The standard TIMIT speech database is used for experimentation. TIMIT database has two datasets, one for training, and another for testing. Along with the TIMIT speech database an additional speech database of isolated utterances and noise has been created by recording through Pratt software in noiseless room environment. Additive white Gaussian noise data is also used for the experiment. A quantity ten of every letter of the English alphabet has been created. The utterances corresponding to the vowels are used from this database. The results of the database created in the room environment are compared with those of TIMIT database.

B. Performance Analysis for the AWGN

The db filter is used for analysis and synthesis. Matlab software version R2010a is used for implementation. The input speech utterances are loaded into Matlab workspace and resampled to 16 KHz in order to increase the sample to sample correlation. The input speech with SNR levels 10dB, 15dB, 20dB, are first subjected to speech versus noise discrimination and are decimated into approximation and detail coefficients by a scaling factor 2 through 2 level WPD wavelet. The details of feature measurements made for speech and silence classification are given in Table I, Table II, Table III and Table IV. The first level subband details are treated as variance, which is used to find universal threshold. The universal threshold is used to find time adaptive threshold. To reduce the impact of noise on the signal thresholding operation is applied to all coefficients i.e., $\omega_{j,m}(k)$, using TAT. A 2-level WPR is applied to thresholded coefficients $\widehat{\omega}_{j,m}(k)$ for estimating the speech. Figure 3 and figure 4 indicate representations of degraded and enhanced speech utterances corresponding to sequence of word utterances "One" respectively in terms of waveform. Finally, Table I, describes the features measured for speech versus noise classification. One can notice significant changes in the measured values of patterns or features of various utterances. This is mainly due to the fundamental property of speech signal being quasiperiodic due to periodic vibration of human vocal chords during the physiological process of speech production and its non stationary property due to changes in the pitch, formant frequencies of human vocal chords. The vocal chord must change its state to produce different speech sounds. The measured patterns are used to distinguish speech from unwanted noise to reduce impact of noise on the speech and hence to improve performance of speech enhancement. Because there is no point in denoising or processing a noisy segment that would not carry any information. Thus separating the noisy or silence portion from the speech increases the performance of speech enhancement and that of automatic speech recognition systems.

TABLE I. PATTERN MEASUREMENTS FOR TIMIT SPEECH UTTERANCES.

TIMIT UTTERANCES	**s101**	**s102**	**s103**	**s104**	**s105**	**s106**	**s107**	**s110**
ZCR	7813	4977	9750	10740	8088	7426	3627	9707
NCC	0.0686	0.3361	0.0827	0.0301	0.1329	0.7706	0.1835	0.0494
ENERGY	3.365	2.0047	2.2929	2.4227	2.8395	2.8871	1.5064	1.8435

TABLE II. PATTERN MEASUREMENTS FOR TIMIT SPEECH UTTERANCES.

TIMIT UTTERANCES	**s201**	**s202**	**s203**	**s204**	**s205**	**s206**	**s207**	**s208**	**s210**
ZCR	10735	7886	16139	9941	20814	10183	8915	9709	9005
NCC	0.0609	0.3738	0.0696	0.1649	0.3388	0.016	0.064	0.6456	0.2394
ENERGY	7.7665	4.8962	6.1545	3.4355	6.2799	4.5669	4.0824	3.1826	7.4224

TABLE III. PATTERN MEASUREMENTS FOR SPEECH UTTERANCES RECORDED IN ROOM ENVIRONMENT.

RECORDED DATA	**A**	**E**	**I**	**O**	**U**
ZCR	448	227	436	442	453
NCC	0.5694	1.0698	0.4947	1.5564	0.3656
ENERGY	205.9218	105.8717	102.0342	214.6707	175.2172

TABLE IV. PATTERN MEASUREMENTS FOR NOISY OR SILENCE SEGMENTS.

BACKGROUND NOISE	**1_s**	**2_s**	**3_s**	**6_s**	**8_s**	**9_s**	**10_s**
ZCR	21048	20729	8329	658	4431	2894	2912
NCC	0.0013	0.001	0.0017	0.0289	0.0064	0.0048	0.0033
ENERGY	6.51E-04	6.70E-04	2.67E-04	0.0818	0.7198	0.6029	0.6611

TABLE V. PATTERN MEASUREMENTS FOR AWGN NOISE RECORDED IN ROOM ENVIRONMENT.

AWGN	N1	N2	N3	N4	N5	N6	N7
ZCR	23299	23537	23607	23577	23596	23785	24060
NCC	2.71E-05	-1.22E-05	-3.25E-05	1.08E-05	1.22E-07	-9.71E-06	1.14E-05
ENERGY	34.043	10.6143	3.3657	1.0628	0.3355	0.1068	0.0337

There is a significant change in the zero crossing rates between speech utterances and that of silence or background noise. The Table I and Table II gives the details of the patterns measured for different set of speech utterances of TIMIT database. The Table III gives the details of the patterns measured for vowels. The Table IV and Table V gives the details of the patterns measured for a set noisy utterances of the database created in room environment and AWGN noise. The figures 3 and figure 4 represent the TIMIT speech sentence and AWGN noise respectively. The wavelet based speech enhancement gives an output SNR of 16dB for speech of input SNR 10dB as this is evident from the figure 5 and figure 6. The experimental results presented in Table I to Table V says that among the patterns measured for speech discrimination from noise, all corresponding patterns are different for speech and noise respectively. Hence all patterns contribute efficiently for efficient classification of speech utterances from the noise. This is mainly due to the fundamental property of speech signals that they posses more energy, high NCC, lesser zero crossings than those of noise or silence.

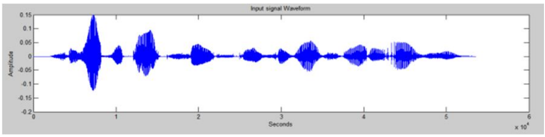

Figure 3. The input TIMIT utterance “she had your dark suit in greasy wash water all year”

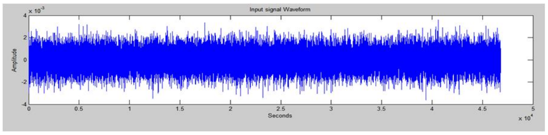

Figure 4. The input AWGN noise.

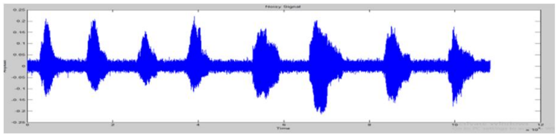

Figure 5. Noisy speech of sequence of word utterance “one” by a female speaker before denoising SNR 10dB

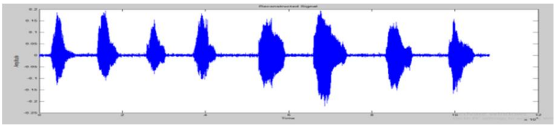

Figure 6. Reconstructed speech of sequence of word utterance “one” by a female speaker after denoising SNR 16dB

V. CONCLUSION

The results of the database created in the room environment are compared with those of TIMIT database. Among the patterns measured for speech discrimination from silence, all the patterns gave significant contributions for efficient classification of speech utterances from the noise. This is due to the fundamental property of speech signals that they have got more energy than that of random noise or silence. The experimental results reveal that the proposed method yields good performance for efficient classification of speech utterances from background noise or silence segments. Thus we conclude that the speech versus silence discrimination step plays a vital role by eliminating unwanted computational complexity. Further the use of SVSD along with the speech enhancement systems, as preprocessing front end in automatic speech recognition systems, further improves their performance by reducing unwanted signal analysis, processing and complexity.

REFERENCES

[1] Patil, Sanjay P., and John N. Gowdy. "Use of baseband phase structure to improve the performance of current speech enhancement algorithms". Speech communication 67 (2015): 78-91.

[2] Sanam and Shahnaz: “A semisoft thresholding method based on Teager energy operation on wavelet packet coefficients for enhancing noisy speech”. EURASIP Journal on Audio, Speech, and Music Processing (2013):25.

[3] Xia B, Bao C. “Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification”. Speech Communication. 2014 May 31;60:13-29.

[4] Sanam, Tahsina Farah, and Celia Shahnaz. "Teager energy operation on wavelet packet coefficients for enhancing noisy speech using a hard thresholding function." signal processing: an international journal 6.2 (2012): 22.

[5] Chen, Shi-Huang, and Jhing-Fa Wang. "Speech enhancement using perceptual wavelet packet decomposition and teager energy operator." Real World Speech Processing. Springer US, 2004. 51-65.

[6] Chang, S. Grace, Bin Yu, and Martin Vetterli. "Adaptive wavelet thresholding for image denoising and compression." IEEE transactions on image processing 9.9 (2000): 1532-1546.

[7] Ephraim, Yariv, and David Malah. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator." IEEE Transactions on Acoustics, Speech, and Signal Processing 32.6 (1984): 1109-1121.

[8] Ephraim, Yariv, and Harry L. Van Trees. "A signal subspace approach for speech enhancement." IEEE Transactions on speech and audio processing 3.4 (1995): 251-266.

[9] D. L. Donoho, "De-noising by soft-thresholding," in *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613-627, May 1995.

[10] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-27, pp. 113–120, Apr. 1979.

[11] Donoho, David L., et al. "Wavelet shrinkage: asymptopia?." Journal of the Royal Statistical Society. Series B (Methodological) (1995): 301-369.

[12] Tavares, R., and R. Coelho. "Speech Enhancement with Nonstationary Acoustic Noise Detection in Time Domain." IEEE Signal Processing Letters 23.1 (2016): 6-10.

[13] J. F. Kaiser, "Some useful properties of Teager's energy operators," *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, Minneapolis, MN, USA, 1993, pp. 149-152 vol.3.

[14] Seok, Jong Won, and Keun Sung Bae. "Speech enhancement with reduction of noise components in the wavelet domain." Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on. Vol. 2. IEEE, 1997.