International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:14, No:1, 2020

Speech Intelligibility Improvement Using Variable Level Decomposition DWT

Samba Raju, Chiluveru, Manoj Tripathy

Abstract—Intelligibility is an essential characteristic of a speech signal, which is used to help in the understanding of information in speech signal. Background noise in the environment can deteriorate the intelligibility of a recorded speech. In this paper, we presented a simple variance subtracted - variable level discrete wavelet transform, which improve the intelligibility of speech. The proposed algorithm does not require an explicit estimation of noise, i.e., prior knowledge of the noise; hence, it is easy to implement, and it reduces the computational burden. The proposed algorithm decides a separate decomposition level for each frame based on signal dominant and dominant noise criteria. The performance of the proposed algorithm is evaluated with speech intelligibility measure (STOI), and results obtained are compared with Universal Discrete Wavelet Transform (DWT) thresholding and Minimum Mean Square Error (MMSE) methods. The experimental results revealed that the proposed scheme outperformed competing methods

Keywords—Discrete Wavelet Transform, speech intelligibility, STOI, standard deviation.

I. Introduction

SPEECH enhancement is the process of extracting a clean speech signal from a noisy speech signal. Variety of conventional methods exist for speech enhancement such as Spectral Subtraction [1], Minimum Mean Square Error (MMSE) [2] and Discrete Wavelet Transform (DWT) denoising [3] etc. All these methods had improved noisy speech quality pretty good, however, improving speech quality does not necessarily improve speech intelligibility [4], and factors explaining why speech quality is not directly linked to speech intelligibility are studied in [5]. Speech Intelligibility (SI) usually refers to a measure for the effectiveness of the understanding of speech. The speech intelligibility has been improved with Computational Auditory Scene Analysis (CASA), which uses an estimated mask for speech enhancement, supervised models are used for estimation of the mask, which improves intelligibility greatly [6]. The problem with the supervised models is that they would require a large training dataset and training time for better generalization; further, these supervised models produce less intelligibility in an unknown environment [7].

DWT is a Multi-Resolution Analysis (MRA), which had been applied to various research areas. Wavelet shrinkage is a simple denoising technique, which performs the thresholding of wavelet coefficients. These coefficients are derived for a particular DWT decomposition level, the threshold value is decided from wavelet coefficients such that it distinguishes the target signal coefficients from the noisy wavelet coefficients. Unfortunately, simple thresholding may not fully separate the target signal from noisy coefficients. Furthermore, under non-stationary noise, the unvoiced speech segments are comparable with noise; hence, it is difficult to distinguish. Wavelet thresholding was uniformly applied to all coefficients which suppress noise along with some unvoiced speech utterances.

The wavelet transform is used to analyze speech signal in time-frequency representation, and the speech signal is typically sparse in time-frequency representation [8]. The proper decomposition level (DL) of DWT can generate sparsed speech and noisy coefficients [9]. In this paper, a novel approach for the detection of DL is presented, which further utilized to improve speech intelligibility. The proposed algorithm is easy to implement and reduces the computational burden, unlike the conventional denoising methods, because, they require prior knowledge of noise SNR. The proposed algorithm is evaluated with speech corpus taken from TIMIT Database [10] and noise corpus took from NOISEX-92 databases [11], and these results are compared with DWT-denoising method and the MMSE algorithm, respectively.

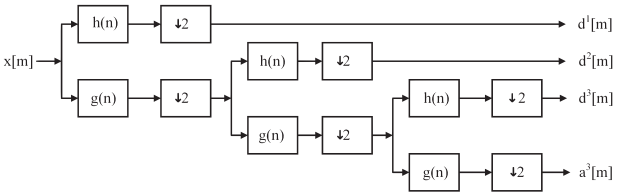

Fig. 1 Three level decomposition structure

II. Speech Enhancement Using DWT

Mallat et al. in [3] show that the DWT is viewed as a multi-stage signal decomposition process using the basic filter bank structure shown in Fig. 1. In this implementation low pass filter $(g(n))$ generate coarse approximation coefficients, and high pass filter $(h(n))$ generate its detail coefficients and redundancy introduced had been removed by using downsampling. The filter bank operates recursively on the low-pass filtered data to generate coarser decompositions of the input signal and its corresponding details.

$$d^j(m) = \sum_{k=0}^{K-1} h(k) d^{j-1}(2m-k) \qquad 0 \le m \le M \quad (1)$$

Samba Raju, Chiluveru and Manoj Tripathy are with the Department of Electrical Engineering, Indian Institute of Technology, Roorkee, 247667 India (e-mail: samba8su@gmail.com).

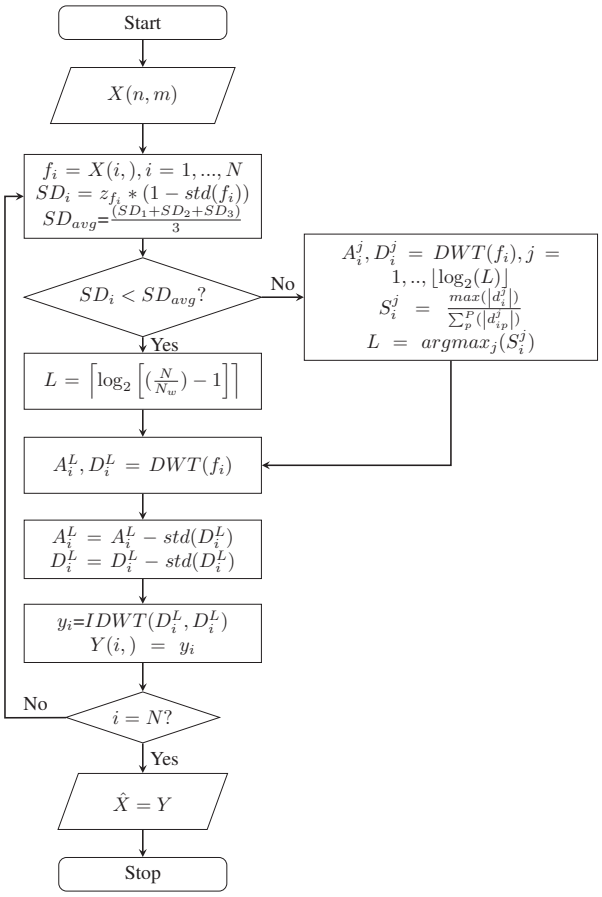

Fig. 2 Flowchart for varible DWT level estimator

$$a^j(m) = \sum_{k=0}^{K-1} g(k) a^{j-1}(2m-k) \qquad 0 \le m \le M \tag{2}$$

$$w_m^j = \left\{ a^j(m), d^j(m) \right\} \tag{3}$$

Donoho and Johnston proposed a denoising method for white noise, which proceeds by thresholding wavelet coeffitients [12], [13], a universal threshold λ is proposed for removing white noise

$$\lambda = \sigma \sqrt{2 \log(M)} \tag{4}$$

with

$$\sigma = MAD/0.6745 \tag{5}$$

where σ is the noise standard deviation. Median absolute deviation (MAD) is estimated in the first scale. The soft thresholding function is defined as

$$T_s(\lambda, w_m) = \begin{cases} sgn(w_m^j)(\left|w_m^j\right| - \lambda) & \text{if } \left|w_m^j\right| \ge \lambda \\ 0 & \text{if } \left|w_m^j\right| < \lambda \end{cases} \tag{6}$$

Finally these wavelet coefficients are used to reconstruct speech signal.

III. Proposed Variable Level DWT Decomposition Denoising

The proposed intelligibility improvement algorithm is based on the sparseness of noisy speech signals in the time-frequency domain. Fig. 2 shows the flow chart of the proposed algorithm, which approximate sparse, noisy speech signal into target and noise coefficients, afterward L^{th} level wavelet coefficients are subtracted with a standard deviation of detail coefficients which improve the intelligibility of speech signal. In the proposed algorithm, the detection of a wavelet decomposition level is crucial because it adequately sparse the noise from the speech signal. In this algorithm, DL is decided with two modules, such as signal dominant and noise dominant modules.

1) Signal Dominant Detector: In this module, we defined a new parameter Speech Detection (SD), for each frame SD value is calculated if this value found to be less than a reference threshold (SD_{ta}) then that frame is called speech dominant frame (SDF) else noise dominant frame (NDF). The reference value is computed as an average of SD for the first three frames of the noisy speech signal. The SD value is presented in equation (7), which is derived with a zero-crossing rate (z) and a standard deviation (σ) for each frame.

$$SD_i = z(\mathbf{x}_i) * (1 - \sigma(\mathbf{x}_i)) \tag{7}$$

Signal dominant detector uses a reference value to differentiate speech signal from noise, the reference value is given by SD_{avg}:

$$SD_{avg} = \frac{SD_1 + SD_2 + SD_3}{3} \tag{8}$$

2) Variable Decomposition Level Detector: If the frame is SDF then DWT-DL is decided in a general way, while if it is an NDF, i.e., the frame has signal mixed with noise or noise alone, then DL has decided with $''peak - to - sum''$ ratio $((S_i^j))$ of detail coefficients. Noise components have less amplitude compared with target signal so S_i^j yields a maximum value for a mixed type of frame and it gives minimum value for same amplitude coefficients (i.e., noise), hence DL is decided with the index (j) which yields maximum value of S_i^j $(L = max_j(S_i^j))$. Equation (9) is used to calculate S_i^j, where, d_i^j represent i^{th} index j^{th} level detail coefficients.

$$S_i^j = \frac{max(\left|d_i^j\right|)}{\sum_p^P(\left|d_{ip}^j\right|)} \tag{9}$$

The overview of proposed algorithm as follows:

1) If the frame is SDF $(t_{avg} \ge t)$ then $L = \left\lceil \log_2 \left[(\frac{N}{N_w}) - 1 \right] \right\rceil$, else,
2) If the frame is NDF $(t_{avg} \le t)$ then L is decided based on the detailed co-efficients
   $L = max_j(S_i^j) = \frac{max(|d_i^j|)}{\sum_m^M(|d_{im}^j|)}$
3) Standard deviation subtracted from wavelet coefficients
   $\hat{A}_i^L(m) = A_i^L(m) - std(D_i^L(m))$
   $\hat{D}_i^L(m) = D_i^L(m) - std(D_i^L(m))$

4) Inverse DWT is applied to reconstruct speech signal $\hat{x}_i(m) = IDWT_j(\hat{A}_i^L(m), \hat{D}_i^L(m))$

The enhanced speech signal is reconstructed using add and overlap method.

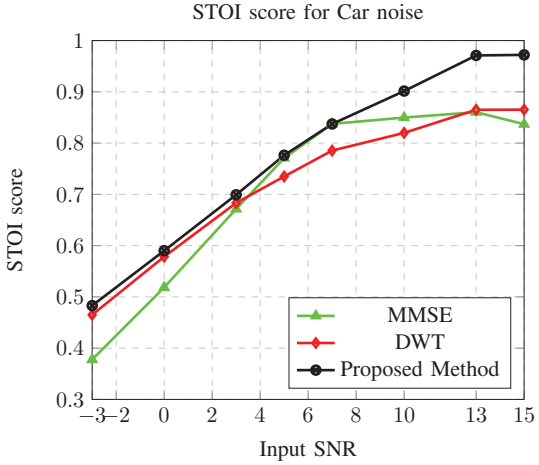

Fig. 3 Experimental results of STOI score for car noise environment

IV. Results and Discussion

The proposed algorithm is evaluated using real-time non-stationary noises like factory, car, babble, and street noise, these noises are considered from the NOISEX-92 dataset. A clean utterance from TIMIT database is corrupted with different noises under different SNR levels (i.e. -3,0,3,5,7,10,13 and 15 dB SNR). The synthetic noisy speech signal is downsampled to 8kHz and a hamming window of width 32 msec with overlapping of 50% is used for framing. The resulted frame has 256 samples, Daubechies mother wavelets with 4^{th} order was considered for DWT.

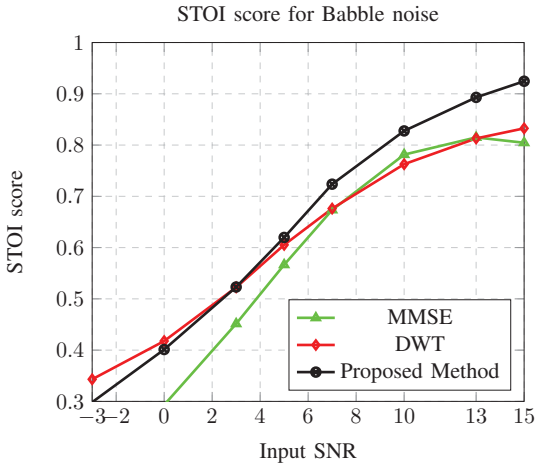

Fig. 4 Experimental results of STOI score for Babble noise environment

Comparative results are drawn using speech intelligibility measure STOI, which showed a better correlation with speech intelligibility compared with other reference objective measure [14]. The proposed algorithm is compared with two conventional methods such as universal thresholding based DWT, and MMSE algorithm. Fig. 3 shows the STOI scores for enhanced speech with car noise, and it is observed that in a low SNR environment, the proposed algorithm shows small improvement whereas, in high SNR level, it shows good improvement. Fig. 4 shows STOI scores for Babble noise, and it is observed that the proposed method dominates the remaining two methods except DWT at -3 and 0 dB SNR levels. Furthermore, it is observed that for the Babble noise environment proposed method gives better results compared to other methods.

TABLE I
STOI Score for Proposed Method, DWT and MMSE

Input SNR	Factory			Street		
	Proposed Method	DWT	MMSE	Proposed Method	DWT	MMSE
-3dB	**0.7878**	0.7227	0.7873	**0.5630**	0.5160	0.5002
0dB	**0.8589**	0.7703	0.8562	0.6140	0.5651	**0.6294**
3dB	0.9090	0.8114	**0.9100**	0.7366	0.6645	**0.7721**
5dB	0.9328	0.8309	**0.9480**	0.7998	0.7292	**0.8278**
7dB	0.9512	0.8487	**0.9616**	**0.8604**	0.7765	0.8600
10dB	**0.9663**	0.8623	0.9159	**0.9082**	0.8114	0.8698
13dB	**0.9476**	0.8526	0.8407	**0.9300**	0.8336	0.8414
15dB	**0.9616**	0.8611	0.8269	**0.9507**	0.8477	0.8248

Table I shows the STOI score for Factory and Street noises, and the proposed method dominates the Conventional DWT method under all SNR levels. MMSE method shows better results at 3, 5 and 7 dB SNR levels for factory noise and 0, 3 and 5 dB SNR levels for street noise, whereas, for remaining SNR levels the proposed method gives good results. Fig. 5 shows the enhanced speech signal for Babble noise using the proposed algorithm, and it is observed that the proposed algorithm is capable of reproducing the original signal without losing its information and therefore, its intelligibility.

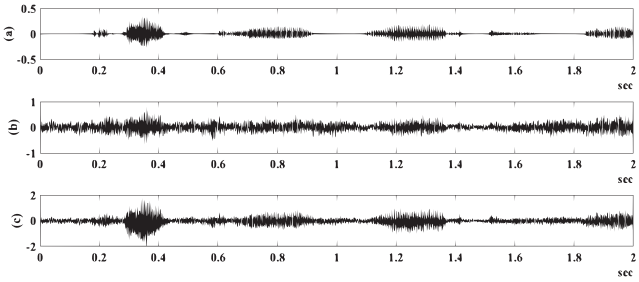

Fig. 5 Speech intelligibility results: (a) Clean speech (b) Noisy speech (Babble noise added at +5dB SNR) (c) Enhanced speech with Proposed Algorithm.

V. Conclusion

The proposed variable level DWT denoising method uses a simple standard deviation for speech intelligibility improvement, and it does not require any prior noise estimation, it is easy to implement, and it decreases computational complexity. The result shows that the intelligibility of noisy speech signal is improved under all noisy environments when compared to DWT-thresholding and MMSE. Especially in a Babble noise environment, the

proposed denoising algorithm shows good improvement in positive SNR levels.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC press, 2007.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a Minimum-Mean Square Error Short-Time Spectral Amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] S. G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[4] G. Kim and P. C. Loizou, "Improving Speech Intelligibility in Noise using Environment-Optimized Algorithms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2080–2090, 2010.

[5] P. C. Loizou and G. Kim, "Reasons Why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, 2010.

[6] D. Wang and J. Chen, "Supervised peech separation based on deep learning: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[7] M. Kolbk, Z.-H. Tan, J. Jensen, M. Kolbk, Z.-H. Tan, and J. Jensen, "Speech Intelligibility Potential of General and Specialized Deep Neural Network based Speech Enhancement Systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[8] S. Y. Low, D. S. Pham, and S. Venkatesh, "Compressive Speech Enhancement," *Speech Communication*, vol. 55, no. 6, pp. 757–768, 2013.

[9] M. Srivastava, C. L. Anderson, and J. H. Freed, "A New Wavelet Denoising Method for Selecting Decomposition Levels and Noise Thresholds," *IEEE Access*, vol. 4, pp. 3862–3877, 2016.

[10] J. S. Garofolo *et al.*, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, pp. 1–6, 1988.

[11] A. Varga and H. J. Steeneken, "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[12] D. L. Donoho and J. M. Johnstone, "Ideal Spatial Adaptation by Wavelet Shrinkage," *biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[13] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

Samba Raju. Chiluveru received his BTech in electronics and communication engineering from JNTU, Hyderabad, India, in 2009, MTech in VLSI and Embedded systems from JNTU, Hyderabad, India, in 2012. He is currently pursuing his PhD from the Department of Electrical Engineering, Indian Institute of Technology, Roorkee, and Uttarakhand, India. His areas of interests are speech enhancement, FPGA, Deep Neural Networks, and digital speech processing.

Manoj Tripathy received his BE degree in electrical engineering from Nagpur University, Nagpur, India, in 1999, MTech. degree in instrumentation and control from Aligarh Muslim University, Aligarh, India, in 2002, and PhD from the Indian Institute of Technology Roorkee, Roorkee, India, in 2008. He is currently working as assistant professor in the Department of Electrical Engineering, Indian Institute of Technology Roorkee, Uttarakhand, India. His fields of interests are wavelets, neural network, optimisation techniques, content-based image retrieval, digital instrumentation, digital protective relays and digital speech processing. Dr. Tripathy is a reviewer for various international journals in the area of power systems and speech.