SINGLE CHANNEL SPEECH ENHANCEMENT BASED ON ZERO PHASE TRANSFORMATION IN REVERBERATED ENVIRONMENTS

Dayana Ribas González, Serguey Crespo Arias, José R. Calvo de Lara

Advanced Technologies Application Center (CENATAV), La Habana, Cuba

ABSTRACT

In this paper we present a single channel speech enhancement proposal in the autocorrelation domain. The main goal of this work is to compensate for speech in reverberated environments. However the proposed method is designed to handle any non-periodic corruption and it can carry out dereverberation and noise reduction at the same time. The work is framed in the REVERB (REverberant Voice Enhancement and Recognition Benchmark) challenge as part of the Speech Enhancement task. Results are given in both, WSJCAM0 Corpus for the simulated data and the MC-WSJ-AV corpus for the real recordings. Four quality measures have been computed for the experiments: cepstral distance, speech-to-reverberation modulation energy ratio, frequency weighted segmental SNR and log-likelihood ratio. Also the computational time of the speech enhancement proposal has been computed by measuring the Wall Clock Time. Promising results have been obtained both for the quality measures and for the computational time. In quality measures, the proposal achieves improvements until 57% using the proposed speech enhancement procedure.

***Index Terms*—** zero phase, autocorrelation, speech enhancement, reverberation, challenge.

1. INTRODUCTION

Reverberation is the collection of reflected sounds from surfaces in an enclosure which distort the structure of a speech signal in both temporal as well as spectral domain. As a result, in the presence of room reverberation the speech intelligibility is affected not only for hearing-impaired and elderly people but also for automatic speech recognition systems (ASR). Room reverberation is one of the major causes of speech degradation and there has been an increasing need for speech dereverberation in speech processing and communication applications. Although it has been studied for decades, speech dereverberation remains a challenging problem both from a theoretical and a practical perspective [1].

Previous works have addressed this issue of reverberation using a variety of methods. Some classification are done based on the number of microphones, in single- and multichannel methods, while others consider the underlying mathematical principles. Source model- or signal features-based speech dereverberation methods estimate the clean speech by using the a priori knowledge about the structure of clean speech signals and how the distortion due to reverberation takes place. Linear-prediction (LP) residual enhancement methods [2], harmonic filtering [3] and speech dereverberation using probabilistic models [4] are typical algorithms in this group. On the other side, speech dereverberation could be carried out from the perspective of signal separation in the cepstral domain [5]. Cepstral liftering [6] and cepstral mean subtraction [7] are the main techniques in this subset. Finally the channel inversion and equalization is another speech dereverberation technique. In this case the idea is to estimate an inverse filter of the acoustic impulse response. However the problem with this technique is that the impulse response must be blindly estimated which is a very challenging task for a single channel case. Minimum mean square error (MMSE) and least-squares algorithms can be used for the estimation.

In the end, all methods have a common objective: to perform an enhancement of the speech signal. The three main axes of speech enhancement are echo cancellation, noise reduction and dereverberation [8]. However in practical applications the speech is often affected by several corruption sources at the same time, for example in applications that acquire samples using mobile devices, such as phones, PDAs and laptops computers and for which hands-free distant talking operation is desirable. In these cases, speech samples could be coexisting with environmental noises, cocktail party effects and reverberation at the same time. In such cases, applying methods exclusively designed for dereverberation is not enough for solving the speech enhancement problem.

In this work, we propose a single channel speech enhancement method using zero phase transformation of the speech signal. Zero phase procedure is a type of transformation in autocorrelation domain, so share its properties. The zero phase sequence energy is concentrated in the origin when the signal spectrum is almost flat, while the sequence is turning periodic when the signal spectrum comes periodic. This behavior helps us to detect the reverberation location in autocorrelation domain and allows us to remove it. Since noise is regularly non-periodic, this method can work for several types of noise, not only for reverberation case. The main advantage of this method is that it is agnostic to the type of corruption affecting the speech, and can handle dereveberation and noise reduction at the same time.

The approach is tested on the *REVERB (REverberant Voice Enhancement and Recognition Benchmark) challenge* [9]. This is a multidisciplinary evaluation part of the *IEEE SPS AASP challenge series*[1] organized by a consortium of research groups to share knowledge among speech researchers for handling reverberant speech. The organizers provide a common evaluation framework (datasets and evaluation metrics) for both tasks: speech enhancement (SE) and ASR techniques in reverberant environments. This framework includes the WSJCAM0 Corpus [10] for the simulated data and the MC-WSJ-AV corpus [11] for the real recordings. The text prompts of the utterances used in RealData and part of the SimData are based on the WSJ 5K corpus [12]. The challenge assumes the scenario of capturing utterances spoken by a single stationary distant-talking speaker with 1-channel (1ch), 2-channel (2ch) or 8-channel (8ch) microphone-arrays in reverberant meeting rooms. The background noise is mostly stationary and the Signal to Noise Ratio (SNR) is modest. Speech enhancement challenge task consists of enhancing noisy reverberant speech with single- or multi-channel speech enhancement techniques and evaluating the enhanced data in terms of objective and subjective evaluation metrics. The ASR challenge task consists of improving the speech recognition accuracy of the same reverberant speech.

This work presents the results of our participation in the speech enhancement task of the challenge. Our proposal needs to be carried out separately for each utterance (utterance-based batch processing) with 1-channel section of the database.

The paper is organized as follow: Section 2 presents the zero phase transformation along with mathematical background and motivation for making the dereberveration in the autocorrelation domain. Section 3 presents the proposed speech enhancement method. In Section 4, the performance of our proposal in the SE challenge task is presented. Also the results for four objective quality measures of the speech and computational cost of the speech enhancement system are presented and discussed. Finally Section 5 provides some conclusions of the work.

2. ZERO PHASE TRANSFORMATION AND REVERBERATION

2.1. Zero phase version of the speech signal

Considering a speech production model with an all pole filter that models the vocal tract. The output of the model is the speech signal $x(n)$ and the resonance frequencies of the vocal tract are the speech formants.

The zero phase version $x_{zp}(n)$ of the signal $x(n)$ is computed by first taking the absolute value of the Fourier Transform representation of the signal and setting the phase to zero (effectively removing it). Let the Fourier Transform [5] coefficient $(X(e^{jw}))$ of $x(n)$ be given by:

$$X(e^{jw}) = |X(e^{jw})| e^{j\angle X(e^{jw})} \tag{1}$$

The inverse Fourier transform of only the signal magnitude then represents the zero phase signal in the time domain. Mathematically, the zero phase version of $x(n)$ is given by:

$$x_{zp}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |X(e^{jw})| e^{jwn} dw \tag{2}$$

2.2. Zero phase transformation and autocorrelation

The zero phase sequence could be generalized as:

$$x_{zp}(n) = IDFT(|X(e^{jw})|^{\beta}) \tag{3}$$

where β is an integer. If $\beta = 2$, we have the autocorrelation function. The zero phase and autocorrelation are very similar transformations, the difference between them is only given by the magnitude power, i.e the β coefficient. Therefore, the frequency response of zero phase will be the square root of the power spectral density of the signal and its dynamic range will be less [13]. However despite the difference in the frequency response, the zero phase has the same properties as the autocorrelation, because zero phase transformation takes place in autocorrelation domain.

In this work our focused is on the autocorrelation property related with periodic signals: *"The autocorrelation function of a periodic signal is also periodic with the same period"* [5]. Thus, the autocorrelation sequence has similar information in each period. Voiced speech is quasi-periodic in a short-time segment, i.e. in a frame. Therefore computing short-time autocorrelation [15] in a frame allows us to obtain a periodic pattern. Fig. 1 shows the $x_{zp}(n)$ of voiced speech segment with 30 frames.

2.3. The reverberation in autocorrelation domain

Fig. 2 shows $x_{zp}(n)$ of a reverberated signal in a segment with 30 frames is represented. The figure shows how the application of the zero phase transformation to a reverberation signal results in a sequence with the energy concentration around the zerolag. Thus the autocorrelation sequence does not have a periodic pattern.

3. THE PROPOSAL

3.1. Speech enhancement in autocorrelation domain

Our proposal is based on the previous work [14]. Where the authors proposed a wide-band noise reduction method using

[1] http://www.signalprocessingsociety.org/technical-committees/list/audio-tc/aasp-challenges/

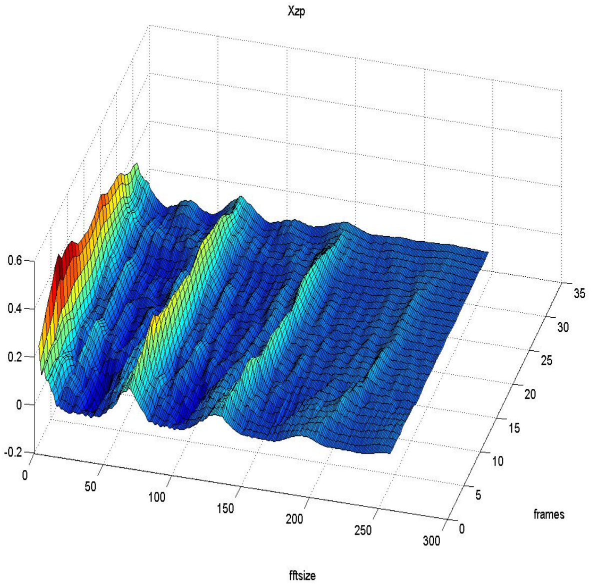

Fig. 1. $x_{zp}(n)$ *of a voiced speech segment with 30 frames.*

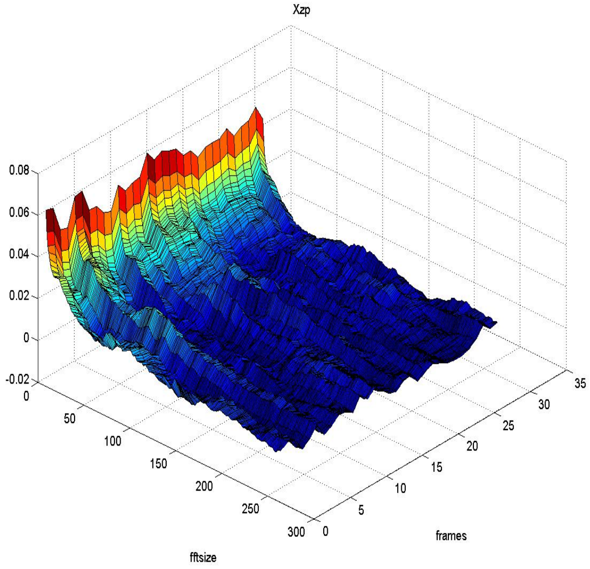

Fig. 2. $x_{zp}(n)$ *of a reverberation signal in a segment with 30 frames.*

zero phase signal, that was tested with some types of stationary and non-stationary noises: tunnel, motor, babble and clap noise.

We are now going to use a similar approach for dereverberation. The idea consists of computing the zero phase version of the signal and once in the autocorrelation domain, to replace the reverberated samples. Then the time sequence is reconstructed from the zero phase estimated sequence returning the enhanced signal. An overview of the method in the following Fig. 3.

3.2. From time domain to autocorrelation domain

First $x_{zp}(n)$ is computed to obtain a representation in the autocorrelation domain with $\beta = 1$. The signal is segmented in frames of 32 ms with an overlap of 10 ms. In order to obtain soft transitions between frames, a Hamming window is applied. Then, the zero phase transformation is implemented by the Fast Fourier Transform (FFT) computation, keeping the magnitude and saving the spectral phase. For the FFT computation, 512 dimensions have been used. Finally the zero phase sequence $x_{zp}(n)$ is obtained by computing the Inverse FFT (IFFT) only with the magnitude.

3.3. Reverberated samples substitution

The reverberated samples substitution is computed for each frame of the $x_{zp}(n)$ sequence. This process is based on:

1. the knowledge of the reverberation location in the autocorrelation domain (section 2.3)
2. the autocorrelation property related with periodic signals (section 2.2)

As was shown in previous section, the reverberation is located around the zerolag, i.e. in the first period of the zero phase sequence. Thus the second period can be used for replacing the corrupted part of the sequence, enhancing the speech frame. From this, we use the period (T_0) of the voiced speech, obtained as the inverse of the pitch. In this approach the pitch is detected using the autocorrelation method [15]. The main peak in the autocorrelation function is at the zerolag location. The location of the next peak gives an estimate of T_0, and the height gives an indication of the periodicity of the signal. Therefore, a peak selection algorithm is used for obtaining a pitch estimation.

Then the samples around zerolag are replaced by the samples around T_0. The amount of reverberated samples define the corrupted segment in the frame, and the number of samples to replace. This number is determined empirically after the processing of a reverberated dataset 10 samples [14]. Fig. 4 presents a schema of the substitution process.

3.4. Time sequence reconstruction

Given the $x_{zp}(n)$ estimation ($x_{zp}est(n)$) in autocorrelation domain, the speech sequence reconstruction basically consists of a process similar to that described in section 3.2. First, the FFT of 512 dimensions is computed obtaining the magnitude of the $x_{zp}est(n)$ sequence. The phase previously saved is coupled with the magnitude, conforming the new spectrogram. With the IFFT the signal is returned to the time do-

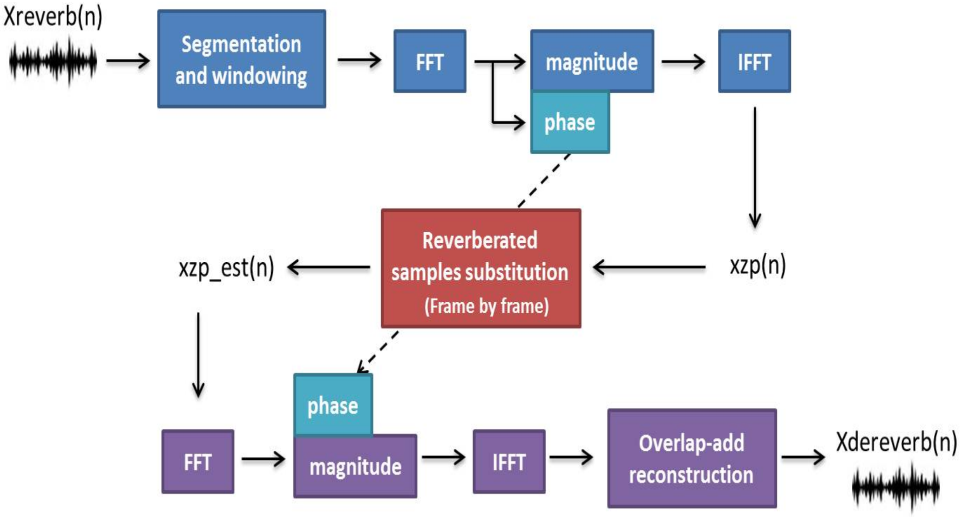
Xreverb(n)
Segmentation and windowing
FFT
magnitude
phase
IFFT
xzp(n)
Reverberated samples substitution (Frame by frame)
xzp_est(n)
FFT
phase
magnitude
IFFT
Overlap-add reconstruction
Xdereverb(n)

Fig. 3. *Overview of general speech enhancement method.*

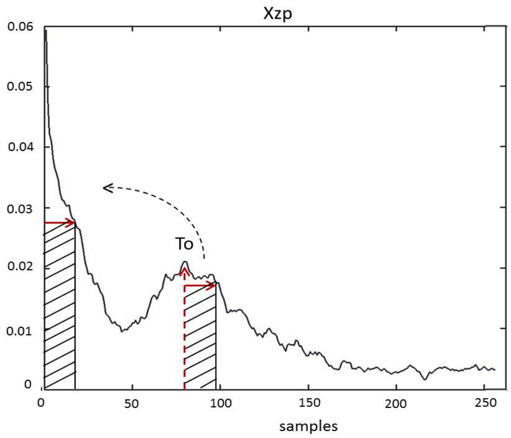

Fig. 4. *Diagram of the reverberated samples substitution process. See text for details.*

main and finally an overlap-add algorithm [16] is performed for converting the separated frames to a temporal sequence again. In this process an inverse Hamming window function is applied for windowing compensation.

4. EXPERIMENTAL SETUP

For performing the SE task we followed the challenge guidelines in [9]. Since this work proposes a single channel method, we use the database recorded with one channel. Evaluation of our proposal is carried out by computing the following objective measures:

- Cepstrum distance (CD) [17]
- Speech-to-reverberation modulation energy ratio (SRMR) [18]
- Log lokelihood ratio (LLR) [19]
- Frequency-weighted segmental Signal to Noise Ratio (FWSegSNR) [19]

CD is a distance between features in the cesptral domain. It is calculated from the observed/enhanced signal and the clean reference. The closer the target feature is to the reference the better is the quality of the signal. Therefore smaller values indicate better speech quality. SRMR measures the energetic relation of the speech and reverberation in the modulation spectral domain. In this case, larger values indicate better speech quality. LLR represents the degree of discrepancy between smoothed spectra of the target and reference signals. It is computed over the Linear Prediction Coefficients (LPC) [20]. Smaller values indicate better speech quality. FWSegSNR is power relation between the speech and noise, computed in the frequency domain. Here, larger values indicate better speech quality.

Also, in order to measure and report the amount of time that the speech enhancement system proposed spends to process the database, the Wall Clock Time (WCT) was computed. This is defined as the amount of time spent in carrying out all the operations other than disk access for read and write, including initialization, zero phase signal computation, enhancement in autocorrelation domain and reconstruction of the speech signal. The WCT was computed in both simulated and recorded data. A reference enhancement code was used to obtain a normalized WCT measure regarding the computer machine features.

4.1. Quality measures results

Table 1 presents the results for the quality measures over the development and evaluation datasets. Shadowed cells highlight those results where our proposal outperforms the baseline. Results for development dataset are a preview of the evaluation dataset results, because the behaviour of the proposed method with respect to the baseline is similar for both datasets. That demonstrates the robustness of the proposed method, indicating that the parameters are not tuned to some specific dataset.

In general, that quality measures reflect that enhanced signals have better quality after the processing with the proposed method as indicated by the LPC-based quality measure. Specially in SRMR and LLR, the proposed method outperforms the baseline for both the development and evaluation dataset. Considering the average of the results in all type of room conditions, our proposal achieves the better performance for all quality measures for LLR, with an improvement of 57%. Also for SRMR and FwSNR the improvements are noticeable too, performing 38.31% and $21, 27\%$ better than the baseline.

On the other hand for cepstral domain, the results are not accurate enough to outperform the baseline results, performing 10.32% worse than the baseline. That is because the enhancement is the result of processing in time domain without consider the bounds of the corrupted segment in the frame, causing aliasing in the spectral domain and consequently affecting the cepstral representation. Future work will be related to developing some method for improving the speech enhancement in the cepstral domain.

4.2. Computational cost results

Table 2 shows the computational cost of our proposed speech enhancement system and the reference enhancement code using the WCT measure. The reference enhancement code was provided by the REVERB challenge [9] organizers equal for all participants, this performs beamforming with two microphone signals. For computing the experiments we used a computer machine with an Intel Core i3-2100 CPU that has 3.10 GHz x 4 processors, a memory capacity of 1.8 GiB and a Linux Ubuntu 13.04, 64bit as operative system. Experiments were executed using Matlab 2013, release A, version: 8.1.0.604.

Table 2. *Computational cost results by the Wall Clock Time.*

Method	SimData	RealData
Reference	268.4141 sec	38.06268 sec
Proposal	251.5522 sec	38.11760 sec

In this measure results are encouraging. In simulated data the proposal performs faster than the reference, computing the enhancement 16.8619 seconds before the reference. On real data the reference performs slightly faster than our proposal, with a difference of 0.05492 seconds. However summarizing the results for all the dataset our proposal performs 16.8069 seconds faster than the reference.

5. CONCLUSIONS

In this work we proposed a single channel speech dereverberation method. Its main advantage is that this method can be used to deal with any non-periodic corruption, making it useful for application in dereverberation and noise reduction at the same time.

For the quality measures in spectral domain results are encouraging, outperforming the baseline almost all the time. However in cepstral domain results are not accurate enough compared to the baseline. In the efficiency measure, despite in real data the reference performs 0.05492 seconds faster than the proposal, then in simulated data the proposal achieves the computation 16.8619 seconds before the reference. Therefore in general the proposal has less computational cost than the reference.

Results are encouraging, however some improvement can be made in the proposed method for obtaining better results. Such as to developing some method for improving the speech enhancement in the cepstral domain. On the other side, in the substitution frame step, the autocorrelation peak selection method for pitch detection was used. However for improving the performance, another more accurate pitch detection algorithm can be applied.

Table 1. *Quality measures results for SE task in development and evaluation test set.*

Quality measures	SimData							RealData		
	Room1		Room2		Room3		Average	Room1		Average
	Near	Far	Near	Far	Near	Far	-	Near	Far	-
Baseline in development test set										
Cepstral distance	1.96	2.65	4.58	5.08	4.20	4.82	3.88	-	-	-
SRMR	4.37	4.63	3.67	2.94	3.66	2.76	3.67	4.06	3.52	3.79
Log-likelihood ratio	0.34	0.38	0.51	0.77	0.65	0.85	0.58	-	-	-
Freq-weighted seg. SNR(dB)	8.10	6.75	3.07	0.53	2.32	0.14	3.48	-	-	-
Proposal in development test set										
Cepstral distance	3.22	3.62	4.58	5.00	4.76	5.04	4.35	-	-	-
SRMR	5.77	6.27	5.18	4.12	4.98	3.93	5.04	5.75	4.90	5.33
Log-likelihood ratio	0.25	0.31	0.36	0.57	0.50	0.66	0.44	-	-	-
Freq-weighted seg. SNR(dB)	6.73	5.10	4.87	2.16	3.96	1.61	4.07	-	-	-
Baseline in evaluation test set										
Cepstral distance	1.99	2.67	4.63	5.21	4.38	4.96	3.97	-	-	-
SRMR	4.50	4.58	3.74	2.97	3.57	2.73	3.68	3.17	3.19	3.18
Log-likelihood ratio	0.35	0.38	0.49	0.75	0.65	0.84	0.58	-	-	-
Freq-weighted seg. SNR(dB)	8.12	6.68	3.35	1.04	2.27	0.24	3.62	-	-	-
Proposal in evaluation test set										
Cepstral distance	3.24	3.59	4.53	5.03	4.76	5.15	4.38	-	-	-
SRMR	6.05	5.98	5.45	4.2	5.01	3.86	5.09	4.78	4.62	4.7
Log-likelihood ratio	0.26	0.31	0.34	0.54	0.5	0.65	0.43	-	-	-
Freq-weighted seg. SNR(dB)	7.13	5.72	5.13	2.74	3.96	1.64	4.39	-	-	-

6. REFERENCES

[1] "Dereverberation. In: Springer Handbook of Speech Processing", Edited by J. Benesty, M. M. Sondhi and Y. Huang, Chap. 46, ISBN 978-3-540-49125-5, Springer-Verlag Heidelberg, pp. 929-943, 2008.

[2] B. Yegnanarayana and P.S Murthy, "Enhancement of reverberant speech using LP residual signal", IEEE Trans. Audio, Speech and Signal Proc., 8(3), pp. 267-281, 2000.

[3] T. Nakatani, M. Miyoshi and K. Kinoshita, "Single microphone blind dereverberation. In: Speech Enhancement", Edited by J. Benesty, S. Makino and J. Chen, Springer Berlin Heidelberg, pp. 247270, Chap. 11, 2005.

[4] H. Attias, J.C. Platt, A. Acero and L. Deng, "Speech denoising and dereverberation using probabilistic models. In: Advances in Neural Information Processing Systems", Edited by S. Thrun, L. Saul, B. Scholkopf, MIT Press Cambridge, Chap. 13, pp. 758764, 2001.

[5] Oppenheim, A.V. and Schafer, R.W, "Discrete-time signal processing", in A.V. Oppenheim [Ed], Prentice-Hall International, Inc., 1989.

[6] B.-H. Juang, L.R. Rabiner and J.G. Wilpon, "On the use of bandpass liftering in speech recognition", Proc. IEEE ICASSP, pp. 765768, 1986.

[7] B.S. Atal: "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", The Journal of Acoustic Society of America, 55(6), pp. 13041312, 1974.

[8] "Speech Dereverberation", Edited by Patrick A. Naylor and Nikolay D. Gaubitch, ISBN: 978-1-84996-055-7, Springer-Verlag London, 2010.

[9] Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Haeb-Umbach, R., Leutnant, V., Sehr, A., Kellermann, W., Maas, R., Gannot, S. and Raj, B., "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech", Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13), 2013.

[10] Robinson, T., Fransen, J., Pye, D., Foote, J. and Renals, S., "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition", Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95), vol.1, pp. 81-84, 1995.

[11] Lincoln, M., McCowan, I., Vepa, J. and Maganti, H.K., "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments", Proceedings of the 2005 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-05), pp. 357-362, 2005.

[12] Paul, Douglas B., Baker and Janet M., "The design for the Wall Street Journal-based CSR corpus" Proceedings of the workshop on Speech and Natural Language (HLT-91), pp. 357-362, 1992.

[13] Mansour, D. and Juang, B.H., "The short-time modified coherence representation and noisy speech recognition", IEEE Trans. Audio, Speech and Signal Proc., 37(6), pp. 795-804, 1989.

[14] W. Thanhikam, Y. Kamamori, A. Kawamura and Y. Iiguni, "Stationary and non-stationary wide-band noise reduction using zero phase signal", IEICE Trans. Fund., vol E95-A (5), 2012.

[15] Rabiner, Lawrence R. and Schafer, Ronald, "Theory and applications of digital speech processing", Prentice Hall, Pearson. ISBN 0-13-603428-4, 2011.

[16] Rabiner, Lawrence R. and Gold, Bernard, "Theory and application of digital signal processing", Englewood Cliffs, N.J.: Prentice-Hall. pp. 6367. ISBN 0-13-914101-4, 1975.

[17] N. Kitawaki, H. Nagabuchi and K. Itoh, "Objective Quality Evaluation for Low-Bit-Rate Speech Coding Systems", IEEE Journal on Selected Areas in Communications, Vol 6(2), 1988.

[18] T. H. Falk, C. Zheng and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech", IEEE T-ASLP, vol 18(7), pp. 1766-1774, 2010.

[19] Loizou, P. C. "Speech Quality Asssessment. In: Multimedia Analysis, Processing and Communications", Edited by D. T. Weisi Lin, Janusz Kacprzyk, Zhu Li, Ebroul Izquierdo, Haohong Wang, Springer, pp. 623-654, 2011.

[20] Atal, B.S and Schroeder, M.R. "Predictive Coding of Speech Signals", Tokyo, Japan: 6th International Congress on Acoustics, 1968.