A Dual Target Neural Network Method for Speech Enhancement

Chang Huai You
Dept. Aural & Language Intelligence
Institute for Infocomm Research
A*STAR, Singapore
email: echyou@i2r.a-star.edu.sg

Minghui Dong
Dept. Aural & Language Intelligence
Institute for Infocomm Research
A*STAR, Singapore
email: mhdong@i2r.a-star.edu.sg

Lei Wang
COLIPS, Singapore
email: wang3stones@outlook.com

***Abstract*—In this paper, we introduce a neural network based speech enhancement method for heavy noisy scenarios. Comprising bidirectional long-short term memory (BLSTM), convolutional neural network (CNN) and deep feedforward neural network (DNN), a neural network architecture with a novel dual target scheme is proposed. The dual target scheme leverages the speech features from both linear and Mel spectral domains in order to retrieve very weak speech spectrum hidden in heavy noise. Experiments show that our proposed neural network architecture outperforms many traditional enhancement methods in terms of objective distortion measure and quality measure.**

Index Terms: speech enhancement, noise reduction, neural network

I. Introduction

Communication along with very annoy environment still causes uncomfortable experience in modern wireless conversation. Reducing much noise from heavy noisy speech while keeping minimal distortion and raising up speech quality still remains an apparent challenge in modern speech signal processing technology [1]. Most of speech-related applications need enhancement for either human listening such as mobile communication and hearing aids [2] [3] [4] or machine recognition like speech/speaker/language recognition [5] [6] [7] [8]).

For many decades, the validity of traditional enhancement method has been proven effectively for stationary and normal-degree noise environment for the purpose of human listening [9] [10] [11] [12]. However, it produces strong artifacts like musical noise and often suffers from considerable distortion for heavy noise situation [13].

Deep neural network modeling has showed the promising results with the supervised training database [14] [15]. Unlike traditional methods which are mathematically derived through statistical signal optimization with certain reasonable assumption, neural network modeling belongs to data driven approach [16]. As a result, selected noise conditions might not be ideal to simulate unknown conditions. One way to overcome the weakness of the trained neural network model is to involve sufficient training conditions and test under widely noisy conditions. Despite of the vulnerability of neural network modeling, deep learning approaches have been proven to be of powerful capability on suppressing both stationary and highly non-stationary noise signals, which is mainly due to its remarkable nonlinear adaptive ability.

Recently, various convolutional neural network (CNN) and recurrent neural network (RNN) based enhancement methods have been developed to achieve significant improvement for speech enhancement for both subjective listening and speech recognition [17] [18] [19]. In [20], the speech enhancer is built by combining one-dimension (1D) CNN and stacked simple recurrent units (SRU) [21], its performance is evaluated for speech denoising and compressed speech restoration.

Principally, CNN is excellent with local filtering, and RNN such as BLSTM [22] is more or less for tracking the sequential information, and forwarding neural network (FNN) is good at simulating the nonlinearity. In order to improve the performance of the neural network model for speech enhancement, the locality and temporal sequential properties of speech as well as the nonlinearity of signal mixture need to be taken into account. However, in most current end-to-end models for speech enhancement, these factors are not fully considered. In this paper, we propose a novel neural network architecture to model the speech enhancement system, where CNN, BLSTM and DNN are properly constructed. Further more, we introduce a dual target cost function to the architecture. The dual target scheme is adopted to generate a reliable loss function from two different perspectives in both Mel-spectral domain and linear spectral domain. The experimental results show the effectiveness of the proposed architecture in terms of signal distortion measure and speech quality measure.

The remainder of the paper is organized as follows. Section II briefly introduces the existing enhancement methods. Section III describes the proposed neural network systems, and section IV reports the performance evaluation. Finally, the conclusion is given in section V.

II. Related Work

A. Traditional Enhancement Method: Wiener Filter

Let $S_k(l)$, $N_k(l)$, and $X_k(l)$ denote the kth spectral component of the clean speech signal, noise, and the observed noisy speech, respectively, where l denotes the time frame corresponding to time t in analysis interval [0, T]. Generally, for spectral domain speech enhancement, the enhanced speech spectrum is given by $\hat{S}_k(l) = G_k(l)X_k(l)$, where $\hat{S}_k(l)$ is the estimate of $S_k(l)$, $G_k(l)$ is the gain function of the enhancement.

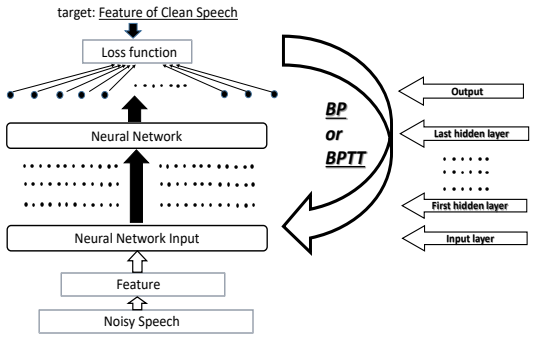

Fig. 1. A general structure of neural network for speech enhancement.

With Gaussian distribution assumption of the respective complex spectra of speech and noise, we have the gain function for Wiener filter as shown below [11]

$$G_k(l) = \frac{\xi_k(l)}{1 + \xi_k(l)} \tag{1}$$

where $\xi_k(l)$ is the *a priori* SNR.

B. Conventional Neural Network Based Enhancement Methods

Speech enhancement can be effectively realized by neural network in both time domain and transform domain due to its powerful nonlinear adaptation ability. Speech denoising system can be modeled by connecting single or different types of neural network in different layers. Fig. 1 shows the general structure of the neural network for speech enhancement. Given a set of input noisy features and the corresponding features from clean speech set as target features, neural network can be trained to model the noise filtering function for speech enhancement via back-propagation (BP) or back-propagation through time (BPTT).

For transform domain modeling, the feature extracted from the speech signal for either input or output target of the neural network can be Mel-frequency cepstral coefficient (MFCC), Mel spectrogram feature, or linear spectrogram feature. Generally, the hidden layers of the neural network can be CNN, RNN or DNN. In [19], the CNN and BLSTM combination is adopted for speech enhancement. In [23], fully CNN is applied for speech enhancement. In [24], DNN is applied into speech enhancement.

III. Proposed Neural Network Architecture

The aim of the study is to do speech enhancement in heavy noise conditions. In real world, the heavy noise condition is the situations in the noisy environment like crowded canteen, crowded traffic station, manufactory working place, window-open driving car and military vehicle. Wireless speech conversation under these conditions are much disturbed and the speech signal are heavily merged in and distorted by background noises.

Fully-connected multilayer FNN, i.e., DNN, is quintessential in deep learning model [14] [15]. Actually, feedforward network can be used to simulate a universal function (e.g., spectral gain function) by adapting the parameters of the function to give approximation of targeted output. However, despite universal function simulation [25], the fully-connected structure of FNNs cannot independently exploit the rich patterns of speech signal. Inspired by the advantages of RNN (especially BLSTM) and CNN, we investigate the network architecture by integrating CNN, BLSTM and DNN to form reliable network for speech enhancement.

Linear spectrogram is short-time Fourier transform (STFT) applied on overlapping segments of the signal. It has been known that human auditory system does not directly perceive frequencies on a linear scale [26], but is good at detecting details in lower frequencies than higher frequencies. Mel spectrogram presents the frequency energy in Mel scale which is close to human auditory property. The Mel scale emphasizes the low frequency area while concentrates speech energy over larger range of high frequency. In fact, from the speech spectrum property, the spectral characteristics is also more like the Mel scale distribution. In other words, more details do distribute in low frequency area (e.g, vowels) while large ambiguous energy distributes over a large range of high frequency (e.g., consonants). By using the Mel-spectrogram, speech spectrum can be separated from noise since background noise does not have the property as speech does, therefore, the weak speech spectral components from noise contamination. However, Mel spectrogram itself has lost much spectral information. It may raise another estimation to approximate the linear spectrogram. An alterative is to train a Mel spectrogram vocoder to convert the Mel spectrogram to linear spectrogram; nevertheless, the distortion from original speech signal is a big problem due to the limitation of the vocoder model. Using linear spectrogram can avoid the problem because the Fourier transform is invertible.

In this paper, we propose a dual target scheme by using two features in the same neural network architecture. The dual target scheme is reflected in a dual cost function that is applied to train the noisy-to-clean enhancement model by using the features in both Mel spectral domain and linear spectral domain. Specifically, with linear spectrogram feature of input noisy speech, we have both Mel spectrogram and linear spectrogram of clean speech signal to be set as target features. Fig. 2 shows the architecture of the proposed training process. First, noisy speech goes through STFT to generate linear spectrogram feature, the feature inputs to a one-layer FNN followed by multi-layer CNN, multi-layer BLSTM, two-layer FNN. Then it flows in two different ways. One way is to one-layer FNN to generate Mel spectrogram feature where we set a Mel spectrogram target for the first loss function. Another way is to merge the Mel feature to form a concatenated vector. Then the concatenated vector goes through DNN (i.e., multi-layer FNN), finally the linear spectrogram is expected to come out from the output layer of the DNN where a linear spectrogram target is set for the second loss function. The final loss cost is the linear combination of the two loss costs

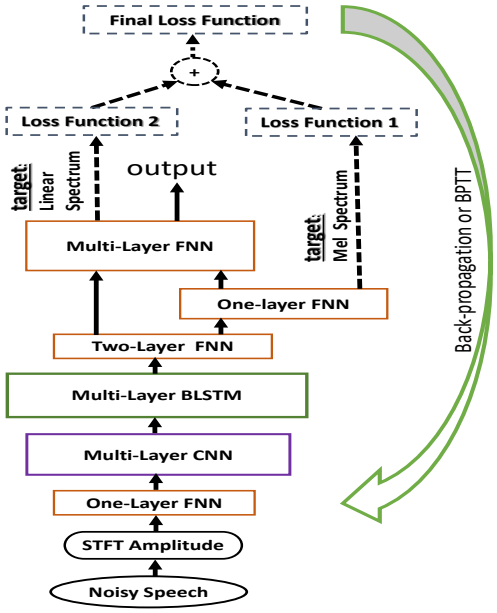

Fig. 2. The proposed neural network architecture with the dual target scheme in consideration of features in Mel spectral domain and linear spectral domain.

as follows:

$$C_{total} = \alpha C_{mel} + (\alpha + 1) C_{linear} \quad (2)$$

where C_{mel} and C_{linear} denote the Mel spectrogram loss cost and linear spectrogram loss cost respectively. α denotes a weight to balance the two loss costs.

IV. Performance Evaluation

We conduct a set of experiments to evaluate the performance over the investigated enhancement methods. Database from Voice Conversion Challenge 2016 (VCC-2016) [27] is selected as clean speech data. There are total 2160 utterances from ten speakers (comprising four male and six female) with 16 kHz sampling rate. We separate 30 utterances for testing and 130 utterances for network validation purpose during model training. The 2000 utterances are used for neural network training purpose. There are 15 types of noise from NOISEX92 [28] and our recorded noise database. 10 types of noises are used for training with ten different segmental signal-to-noise ratios (segSNRs). The ten segSNRs are -10 dB, -7dB, -5dB, -2 dB, 0 dB, 2 dB, 5 dB, 10 dB, 20 dB and clean speech respectively. Thus, we have 200,000 pairs of noisy-to-clean speech utterances for training; and similar 13,000 pairs of utterances for validation. For testing, we use the different 5 types of noise, and three very low segmental SNRs are mainly evaluated: -5 dB, 0 dB, 5 dB. Thus, there are 450 noisy utterances for testing.

We setup typical baseline systems and the proposed systems for comparison as follows:

- **Wiener** [11]: Traditional Wiener filter.
- **3DNN2048** [29]: A deep neural network (DNN) contains three hidden layers with 2048 neurons in each layer, but instead of symmetric context window, it uses causal context window of size 7.
- **CNN1d-BLSTM**: it simulates **CNN2d-BLSTM** but with 1D CNN containing 256 kernels of size 5 with stride 1. In order to fair comparison with our other reference system, the two BLSTMs have only 256 neurons in each layer. One layer FNN is used as the output layer.
- **CMP:1:Single**: Using the proposed neural network architecture as indicated in Figure 2 with single target scheme, i.e. only linear-spectrum target cost applied, where the number of layers in Multi-Layer CNN = 2, the number of layers in Multi-Layer BLSTM = 2, and the number of layers in Multi-Layer FNN = 3.
- **PRO:1:Dual**: Proposed neural network architecture as indicated in Figure 2 with dual target scheme, where the number of layers in Multi-Layer CNN = 2, the number of layers in Multi-Layer BLSTM = 2, and the number of layers in Multi-Layer FNN = 3.
- **CMP:2:Single**: Using the proposed neural network architecture as indicated in Figure 2 with single target scheme, i.e. only linear-spectrum target cost applied, where the number of layers in Multi-Layer CNN = 2, the number of layers in Multi-Layer BLSTM = 2, and the number of layers in Multi-Layer FNN = 5.
- **PRO:2:Dual**: Proposed neural network architecture as indicated in Figure 2 with dual-target scheme, where the number of layers in Multi-Layer CNN = 2, the number of layers in Multi-Layer BLSTM = 2, and the number of layers in Multi-Layer FNN = 5.

TABLE I
Speech distortion measure for the investigated enhancement methods in three different SNRs.

Distort. Measure	MBSD			IS Distortion		
Enhancer\SegSNR	**-5 dB**	**0 dB**	**5 dB**	**-5 dB dB**	**0 dB**	**5**
Noisy	4.104	2.473	1.455	2.094	1.515	1.031
Wiener	1.371	0.792	0.564	4.810	3.942	3.056
3DNN2048	3.173	1.627	0.904	1.940	1.248	0.928
CNN1d-BLSTM	1.016	0.623	0.501	1.491	1.206	0.927
CMP:1:Single	1.045	0.569	0.432	1.502	1.187	0.904
PRO:1:Dual	0.904	0.525	0.407	1.490	1.159	0.896
CMP:2:Single	0.972	0.536	0.413	1.496	1.163	0.901
PRO:2:Dual	0.829	0.506	0.357	1.266	0.903	0.729

TABLE II
Speech quality measure for the investigated enhancement methods in three different SNRs.

Distort. Measure	SegSNR (dB)			PESQ		
Enhancer\SegSNR	**-5 dB**	**0 dB**	**5 dB**	**-5 dB**	**0 dB**	**5 dB**
Noisy	-5.00	0.00	5.00	1.644	2.023	2.400
Wiener	0.89	3.85	7.06	2.121	2.483	2.755
3DNN2048	-3.94	1.59	3.78	1.773	2.206	2.576
CNN1d-BLSTM	0.21	2.80	6.82	2.186	2.473	2.681
CMP:1:Single	0.90	3.60	6.90	2.245	2.630	2.825
PRO:1:Dual	1.14	3.89	7.57	2.386	2.713	2.932
CMP:2:Single	1.03	3.74	7.28	2.312	2.667	2.872
PRO:2:Dual	1.32	4.21	7.79	2.401	2.809	2.960

In speech enhancement, 1D CNN may be more efficient than and 2D CNN [23]. The statement is somewhat consistent with the statistic independent assumption of speech spectral com-

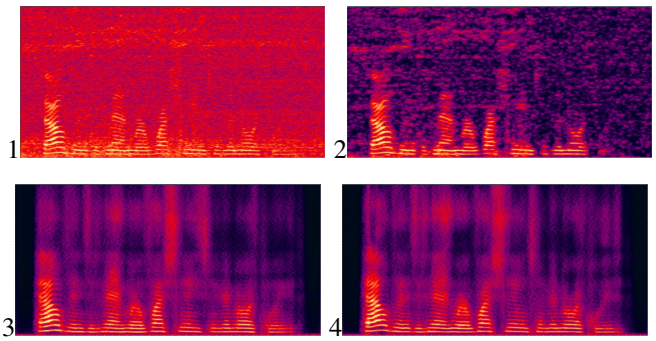

Fig. 3. Spectrogram comparison for noisy with -5 dB segmental SNR contaminated by unknown *Buccaneer2* noise from NOISEX-92 [28], the estimated spectrograms from five different enhancement methods. 1) **Noisy speech**; 2) **Wiener**; 3) **CNN1d-BLSTM**; 4) **PRO:2:Dual**.

ponents in the traditional spectral domain speech enhancement algorithms [3] [10] [4]. In our proposed system, 1D CNN is selected for the two-layer CNN with 256 kernels of size 5 and stride of size 1 for each layer. The number of neurons in each of the two-layer BLSTM is set to 256. And number of neurons in each layer of the multi-layer FNN is 256. Dimension of Mel spectrum is set to 80, while the dimension of the linear spectral amplitude is 513. Batch normalization is used to accelerate model training and drop out is applied in each CNN output. Most of the hidden layers are with rectified linear unit (ReLu) as activation function. For neural network parameter update, the learning rate is set to 0.001 and weight decay is set to 10^{-6}, and gradient clip is applied with threshold set to 1.0. The batch size is set to 16 utterances (each utterance length is in a range of between 1 to 10 seconds) for each iteration. The loss function is based on mean squared error criterion.

Table I shows the distortion measurement in terms of Itakura-Saito (IS) distortion [31] and modified Bark spectral distortion (MBSD) [30]. Table II shows the speech quality comparison in terms of signal-to-noise ratio and perceptual evaluation of speech quality (PESQ) [32]. It can be seen that our proposed neural network architecture outperforms almost all of the enhancers in terms of the four objective measurements. In Fig. 3, four different enhancers are selected for spectrogram comparison for -5 dB heavy noise condition. Wiener filter gives much noise suppression but with apparent musical noise in the background. Our proposed **PRO:2:Dual** shows quite clear spectral outline as compared to the others.

In order to investigate the effectiveness of the dual target against the single target, we show the experimental result for the same proposed neural network architecture with and without Mel-spectral cost function. It is done by introducing the single target **CMP:1:Single** and **CMP:2:Single** for comparison. From the experimental result, it can be observed that the dual target based **PRO:1:Dual** and **PRO:2:Dual** give better performance than single target based **CMP:1:Single** and **CMP:2:Single** in terms of MBSD, IS, segSNR and PESQ.

V. Conclusion

In this paper, a neural network architecture comprising CNN, BLSTM and DNN is investigated, and a dual target scheme is proposed. The aim of the dual target scheme is to retrieve weak speech spectral component from heavy noise. The dual target combining linear spectrogram and Mel-spectrogram increases the speech information by leveraging different aspects of speech feature representation. Performance evaluation shows the validity of the proposed neural network architecture.

VI. Acknowledgements

The authors would like to thank Ms. Aw Ai-Ti for her valuable comment on the neural network architecture design.

References

[1] V. Stahl, A. Fisher and R. Bippus, "Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering," *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, ICASSP, 2000.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109-1121, Dec. 1984.

[4] C.H. You, S.N. Koh, and S. Rahardja, "β-Order MMSE Spectral Amplitude Estimation for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 4, pp. 475-486, Jul. 2005.

[5] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, A. Acero, "Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor," *IEEE Trans. on Audoi, Speech, and Language Process.*, vol. 15, no. 5, July 2008.

[6] C. H. You and B. Ma, "Spectral-Domain Speech Enhancement for Speech Recognition," *Speech Comm.*, ol. 94, Aug. 2017.

[7] R. Tong, B. Ma, K-A Lee, C. H. You, D. Zhu, T. Kinnunen, H. W. Sun, M. H. Dong, E-S Chng, H. Li, "Fusion of acoustic and tokenization features for speaker recognition," *Chinese Spoken Language Processing: 5th International Symposium*, ISCSLP 2006.

[8] Y. Shi, Q. Huang, and T. Hain, "Robust Speaker Recognition Using Speech Enhancement And Attention Model," *Odyssey 2020 The Speaker and Language Recognition Workshop*, Nov. 2020.

[9] C.H. You, S.N. Koh, S. Rahardja Subband Kalman filtering incorporating masking properties for noisy speech signal *Speech Communication*, 49 (7-8), 558-573, 2007.

[10] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-33, No. 2, pp. 443-445, Apr.1985.

[11] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Processing*, vol. 10, pp. 1043–1051, 2003.

[12] P. C Loizou, "Speech enhancement: theory and practice," CRC press, 2013.

[13] O. Cappé, "Elimination of The Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, pp. 345-349, 1994.

[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing* (TASLP), vol. 23, no. 1, pp. 7–19, 2015.

[16] X. Lu, Y. Tsao, S. Matsuda and C. Hori, "Speech enhancement based on deep denoising autoencoder," *In Proc. Interspeech*, pp. 436–440, 2013.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," ArXiv, vol. abs/1505.04597, 2015.

[18] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *ISMIR*, 2018.

[19] H. Zhao, S. Zarar, I. Tashev and C-H. Lee "Convolutional-Recurrent Neural Networks for Speech Enhancement," *IEEE Int. Conf. Acoustics Speech and Signal Processing*, ICASSP, April 2018.

[20] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, "WaveCRN: An Efficient Convolutional Recurrent Neural Network for End-to-end Speech Enhancement," *IEEE Signal Processing Letters*, April 2020.

[21] X. Cui, Z. Chen, and F. Yin, "Speech enhancement based on simple recurrent unit network," Applied Acoustics, vol. 157, pp. 107019, 2020.

[22] S. Hochreiter and J. Schmidhuber, "Long shortterm memory," *Neural computation*, vol. 9, no. 8, pp. 735–1780, 1997.

[23] S. R. Park and J. W. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," *Interspeech*, pp. 1993-1997, Aug. 2017

[24] X. Lu, Y. Tsao, S. Matsuda, C. Hori "Speech Enhancement Based on Deep Denoising Autoencoder," *INTERSPEECH*, 2013

[25] Kurt Hornik, Maxwell Stinchcombe, and Halbert White, "Multilayer feedforward networks are universal approximators," Neural networks, vol. 2, no. 5, pp. 359–366, 1989.

[26] C. H. You, S. N. Koh, and S. Rahardja, "Masking-Based β-Order MMSE Speech Enhancement", *Speech Communication*, Vol. 48, Issue 1, pp. 57-70, Jan. 2006.

[27] http://www.vc-challenge.org/vcc2016/

[28] A. Varga and H. 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, Jul. 1993.

[29] S. Mirsamadi and I. Tashev, "Causal speech enhancement combining data-driven learning and suppression rule estimation.," in INTERSPEECH, 2016, pp. 2870–2874

[30] W. Yang, M. Dixon, and R. E. Yantorno, "Modified bark spectral distortion measure which uses noise masking threshold," *IEEE Workshop on Speech Coding for Telecommunications Proceedings*, pp. 55-56, Oct. 1997.

[31] S.R. Quackenbush, T.P. Barnwell and M.A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[32] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), vol. 2, pp. 749–752 vol.2, 2001.