

# Speech Enhancement Based on Reducing the Detail Portion of Speech Spectrograms in Modulation Domain via Discrete Wavelet Transform

Shih-kuang Lee<sup>1</sup>, Syu-Siang Wang<sup>2</sup>, Yu Tsao<sup>2</sup>, Jieh-weih Hung<sup>1</sup>

<sup>1</sup>National Chi Nan University, Taiwan

<sup>2</sup>Academia Sinica, Taiwan

s105323501@mail1.ncnu.edu.tw, sypdbhee@gmail.com, yu.tsao@citi.sinica.edu.tw, jwhung@ncnu.edu.tw

## Abstract

In this paper, we propose a novel speech enhancement (SE) method by exploiting the discrete wavelet transform (DWT). This new method reduces the amount of fast time-varying portion, viz. the DWT-wise detail component, in the spectrogram of speech signals so as to highlight the speech-dominant component and achieves better speech quality. A particularity of this new method is that it is completely unsupervised and requires no prior information about the clean speech and noise in the processed utterance. The presented DWT-based SE method with various scaling factors for the detail part is evaluated with a subset of Aurora-2 database, and the PESQ metric is used to indicate the quality of processed speech signals. The preliminary results show that the processed speech signals reveal a higher PESQ score in comparison with the original counterparts. Furthermore, we show that this method can still enhance the signal by totally discarding the detail part (setting the respective scaling factor to zero), revealing that the spectrogram can be down-sampled and thus compressed without the cost of lowered quality. In addition, we integrate this new method with conventional speech enhancement algorithms, including spectral subtraction, Wiener filtering, and spectral MMSE estimation, and show that the resulting integration behaves better than the respective component method. As a result, this new method is quite effective in improving the speech quality and well additive to the other SE methods.

**Index Terms:** discrete wavelet transform, speech enhancement, spectrogram, noise reduction, temporal processing

## 1. Introduction

In various speech-related applications, an input speech signal often suffers from environmental noise and requires further processing with a speech enhancement method to improve the associated quality before being of use. By and large, speech enhancement methods can be divided into two classes: unsupervised and supervised. Unsupervised methods, such as spectral subtraction (SS) [1, 2, 3], Wiener filtering [4, 5], short-time spectral amplitude (STSA) estimation [6] and short-time log-spectral amplitude estimation (logSTSA) [7]. By contrast, supervised speech enhancement methods use a training set to learn distinct models for clean speech and noise signals, which examples include codebook-based approaches [8] and hidden Markov model (HMM) based methods [9].

Conventional speech enhancement methods usually process a noisy utterance in a frame-wise manner, viz. to enhance each short-time period of the utterance nearly independently. However, recent researches show that considering the inter-frame variation over a relatively long span of time can contribute

to superior performance in speech enhancement. Some well-known methods along this direction include modulation domain spectral subtraction [10], modulation-domain Wiener filtering and Kalman filtering [11, 12]. In addition, compared with the conventional Fourier transform in which only the frequency parts are considered, the discrete wavelet transform (DWT) [13] takes care of both the time and frequency aspects of a signal and is becoming popular in speech analysis. For example, the well-known wavelet thresholding denoising (WTD) [14] uses the wavelet transform to split the time-domain signal into subbands and then performs thresholding. A recent research [15] applies the DWT to the plain speech feature time series and simply keeps the obtained approximation portion, which achieves data compression and noise robustness in recognition simultaneously.

Partially inspired by the aforementioned ideas, in this study we propose to employ the discrete wavelet transform to analyze the spectrogram of a noisy utterance along the temporal axis, and then devalue the resulting detail portion with a hope to reduce noise effect in order to promote speech quality. In spite of the simplicity of its implementation, the preliminary evaluation results indicate that the proposed method can provide input signals with better perceptual quality. It is also shown that this new method can be paired with several well-known speech enhancement methods to achieve even better performance.

The remainder of this study is organized as follows: Section 2 gives the detailed procedure of the newly proposed method together with the associated discussions and a preliminary test using an example utterance. The experimental setup is described in Section 3, and Section 4 includes the detailed experimental results and the corresponding analyses. Finally, a brief concluding remark including the future avenues is stated in Section 5.

## 2. Proposed method

Here, we present a novel speech enhancement method that employs the DWT to process the spectrogram of a speech signal. This method has a flowchart shown in Figure 1 and consists of the following four steps:

- Step 1: Create the spectrogram  $\{X[m, k], 0 \leq m \leq L - 1, 0 \leq k \leq K - 1\}$  for a given time-domain signal  $x[n]$ , where  $m$  and  $k$  are respectively the indices of frame and acoustic frequency, and  $L$  and  $K$  are the total numbers of frames and acoustic frequency points, respectively.
- Step 2: Use a one-level DWT to decompose the magnitude spectral sequence  $\{|X[m, k]|, 0 \leq m \leq L - 1\}$  with respect to any specific acoustic frequency index  $k$ . That is, we apply the DWT along with the horizontal axis of the

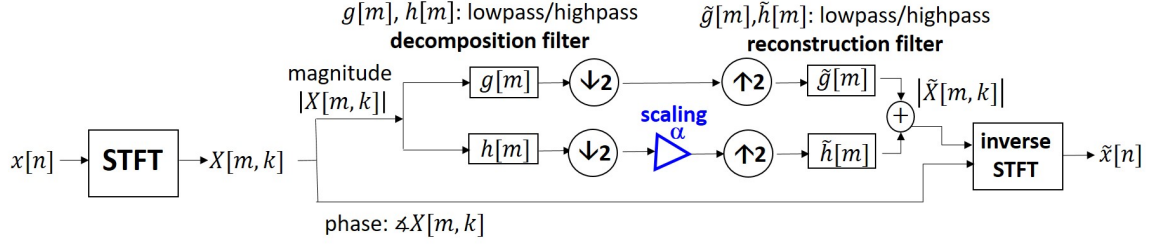


Figure 1: *The flowchart of ModWD.*

spectrogram. The output of the one-level DWT consists of the approximation part  $\{X_a[m, k]\}$  and the detail part  $\{X_d[m, k]\}$ , both of which have approximately half the length of the input  $\{X[m, k]\}$  due to the factor-2 down-sampling.

Step 3: Reduce the detail part  $\{X_d[m, k]\}$  by multiplying a factor  $\alpha$  less than 1 while keeping the approximation part  $\{X_a[m, k]\}$  unchanged for the subsequent one-level inverse DWT (IDWT). That is, we feed the original approximation part and scaled detail part into the IDWT to reconstruct the magnitude spectral sequence  $\{|\hat{X}[m, k]|\}$  associated with the acoustic frequency index  $k$ .

Step 4: The new time-domain signal  $\hat{x}[n]$  is created by applying the inverse STFT to the updated spectrogram, which consists of the new magnitude spectrogram  $\{|\hat{X}[m, k]|\}$  and the original phase spectrogram  $\{\angle X[m, k]\}$ .

The main concepts and characteristics of this new method are as follows:

1. The DWT applied to the temporal sequence of magnitude spectrum is analogous to separating it into two sets of different modulation frequencies, each of which corresponds to a distinct rate of change in the temporal axis. Reducing the detail portion of the DWT output amounts to emphasizing the low modulation-frequency part of the original magnitude spectral sequence, and thus this new method is termed by the modulation-domain wavelet denoising, with a short-hand notation ModWD.
2. According to [16], for a speech signal the linguistic information is mainly located at the modulation frequencies within the range 1 – 16 Hz, with a dominant modulation frequency of 4 Hz. A plenty of research reports [17, 18, 19, 20, 21] also show that highlighting the relatively low modulation frequency components in the speech feature time series benefits the speech quality and recognition accuracy significantly in an adverse environment. As a result, we believe that alleviating the detail portion while preserving the approximation portion of the magnitude spectral sequence can further enhance the speech component and reduce noise effect.
3. In addition to sub-band separation, another operation of the DWT is a factor-2 down-sampling. The newly proposed ModWD simply keeps the factor-2 down-sampled low-pass filtered sequence (the approximation portion) for the subsequent processing. When applying ModWD in a client-server transmission system, the operations of creating a spectrogram, one-level DWT and discarding the detail portion can be implemented on the client side,

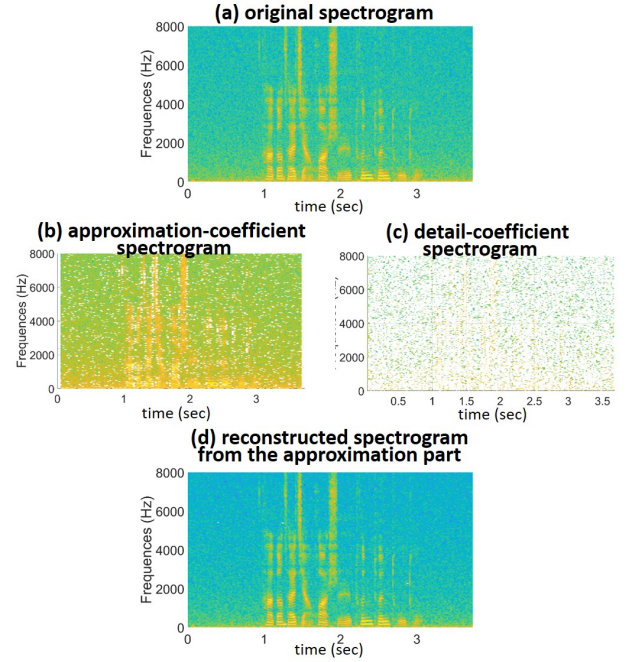


Figure 2: *The different forms of spectrograms of an utterance in the process of ModWD with  $\alpha = 0$ : (a) original spectrogram (b) approximation-coefficient spectrogram (c) detail-coefficient spectrogram (d) reconstructed spectrogram*

and the data being transmitted to the server side for further IDWT and reconstructing a time-domain signal is just half the size of the original spectrogram. Therefore, the presented ModWD is expected to improve the transmission efficiency without the cost of losing predominant speech information.

Figures 2(a)(b)(c)(d) depict the various types of magnitude spectrograms of an utterance during the process of ModWD. Comparing Figures 2(a)(b)(c), the approximation-coefficient spectrogram is shown to contain much more correlation with the original spectrogram than the detail-coefficient spectrogram. Moreover, from Figures 2(a)(d), the new spectrogram reconstructed from the approximation-coefficient spectrogram alone is quite close to the original spectrogram. Therefore, from these figures we provide a preliminary confirmation for the effectiveness of the proposed ModWD that highlights the speech-dominant component in speech purely depending on the DWT-wise approximation portion.

### 3. Experimental Setup

This section presents the database, configurations of the speech enhancement systems, and the evaluation metric.

#### 3.1. Speech data preparation

The experiments use the utterances included in the Aurora-2 database[22], which contains connected English digit utterances generated by both female and male speakers at a sampling rate of 8 kHz. Parts of these utterances are contaminated by various types of noise at different SNRs. In the experiments, 50 airport-noise corrupted utterances belonging to a single speaker were used to form the test set. The SNR levels of the noise-corrupted utterances were varied from 0 dB to 20 dB, with a step of 5 dB.

#### 3.2. Speech enhancement setup

Some information about the setup of ModWD used in this study is as follows:

- Each utterance was split into overlapped frames. The frame duration and frame shift were set to 20 ms and 10 ms, respectively. A Hamming window was then applied to each frame signal.
- The number of frequency bins for the short-time Fourier transform (STFT) was set to 256.
- The biorthogonal 3.7 wavelet basis was used for the DWT and inverse DWT of ModWD.
- The scaling factor  $\alpha$  used in ModWD was set to 0, 0.25, 0.5 and 0.75.

#### 3.3. Objective evaluation metric

Perceptual estimation of speech quality (PESQ) [23] was used as the evaluation metric. PESQ indicates the quality difference between the enhanced and clean speech signals, and it is analogous to the mean opinion score, which is a subjective evaluation index. The PESQ score ranges from 0.5 to 4.5, and a high score indicates that the enhanced utterance is close to the clean utterance.

## 4. Experimental results and discussions

We first evaluated the proposed ModWD in its capability of enhancing noisy utterances. Then ModWD was integrated with several well-known speech enhancement methods to see if further improvement of speech quality could be achieved.

#### 4.1. ModWD with various settings of the scaling factor $\alpha$

Table 1 lists the PESQ results with respect to the tested utterances for the baseline and the counterparts processed by ModWD at different values of the scaling factor  $\alpha$  for the detail portion. (Notably the unprocessed baseline is identical to ModWD with  $\alpha = 1$ ). From this table, some observations can be made:

1. For all tested utterances, the associated PESQ score always gets lower as the SNR becomes worse, indicating that PESQ is an appropriate metric to reflect the degree of noisy distortion in speech.
2. Compared with the baseline results, the proposed ModWD with the scaling factor  $\alpha$  less than 1 can achieve higher PESQ results for almost all tested utterances, with the only exception being ModWD with  $\alpha = 0$  at 0 dB

Table 1: PESQ results for ModWD with different assignments of the scaling factor  $\alpha$

SNR	0	5	10	15	20	Avg.
baseline	1.300	1.768	2.060	2.391	2.780	2.060
$\alpha = 0.75$	1.307	1.779	2.070	2.404	2.789	2.070
$\alpha = 0.50$	1.319	1.788	2.078	2.415	2.795	2.079
$\alpha = 0.25$	1.322	1.794	2.083	2.422	2.797	2.084
$\alpha = 0$	1.289	1.798	2.086	2.428	2.797	2.079

SNR. Therefore, the potential of reducing noise effect of the proposed ModWD is clearly revealed. As an aside, the possible explanation for the degraded performance of ModWD with  $\alpha = 0$  at 0 dB SNR is that significant extra distortion is introduced to a noisy spectrogram by completely discarding the associated detail (high-pass) portion.

3. Setting the factor  $\alpha$  to be either 0.25 or 0 in ModWD can give the best possible performance for the cases of SNR greater than 0 dB, which agrees with the findings in the past research [16] that high modulation frequency components in speech contain less linguistic information and are vulnerable to noise. Substantially reducing or even discarding these components can benefit the speech quality by reducing noise without the expense of introducing significant speech distortion.
4. ModWD with  $\alpha = 0$  indicates that only the DWT-wise approximation portion is required to participate in the subsequent inverse DWT since the detail portion is totally zeroed out. In this case, ModWD is computationally efficient for implementation and can achieve higher transmission efficiency in a client-server architecture because only the factor-2 down-sampled approximation portion has to be sent.

#### 4.2. Cascading ModWD and other speech enhancement methods

Next, we integrated ModWD with several well-known speech enhancement algorithms to see whether such integration can further improve the quality of the testing utterances compared with the individual component method. The algorithms to be integrated include multi-band spectral subtraction (SS) [3], Wiener filtering (WF) [5], short-time log-spectral amplitude estimation (STSA) [6] and short-time log-spectral amplitude estimation (logSTSA) [7]. Please note that interchanging the cascading order of any two algorithms discussed here will behave differently since they are non-linear operations. As for the notations used later, both “ $A - B$ ” and “ $B - A$ ” are the cascade of methods  $A$  and  $B$ , while “ $A - B$ ” indicates performing method  $A$  first and then method  $B$ , and “ $B - A$ ” goes the other way around.

The PESQ results for the integration of ModWD and any of SS, WF, STSA and logSTSA are listed in Tables 2 and 3. Here the scaling factor  $\alpha$  of ModWD is set to either of 0 and 0.25. Figures 3 and 4 further summarize these PESQ values averaged over different SNR cases for ease of comparison. From these tables and figures, we have the following findings:

1. In terms of the PESQ values averaged over the five SNR cases achieved by each individual method, logSTSA behaves the best, followed by WF, STSA, SS and ModWD in turn. It is not surprising that the improvement brought

Table 2: PESQ results for various SE methods including ModWD with  $\alpha = 0$

SNR	0	5	10	15	20	Avg.
ModWD	1.289	1.798	2.086	2.428	2.797	2.079
SS	1.489	2.025	2.341	2.668	3.007	2.306
WF	1.723	2.126	2.412	2.726	2.983	2.394
STSA	1.700	2.141	2.430	2.724	2.974	2.393
logSTSA	1.840	2.234	2.523	2.802	3.058	2.491
ModWD -SS	1.479	2.029	2.366	2.688	3.014	2.315
SS -ModWD	1.500	2.048	2.317	2.694	2.998	2.311
ModWD -WF	1.676	2.079	2.409	2.720	2.967	2.370
WF -ModWD	1.772	2.133	2.441	2.737	2.963	2.409
ModWD -STSA	1.696	2.156	2.449	2.749	2.993	2.409
STSA -ModWD	1.735	2.169	2.445	2.741	2.977	2.413
ModWD -logSTSA	1.813	2.227	2.520	2.819	3.076	2.491
logSTSA -ModWD	1.842	2.240	2.519	2.810	3.054	2.493

Table 3: PESQ results for various SE methods including ModWD with  $\alpha = 0.25$

SNR	0	5	10	15	20	Avg.
ModWD	1.322	1.794	2.083	2.422	2.797	2.084
SS	1.489	2.025	2.341	2.668	3.007	2.306
WF	1.723	2.126	2.412	2.726	2.983	2.394
STSA	1.700	2.141	2.430	2.724	2.974	2.393
logSTSA	1.840	2.234	2.523	2.802	3.058	2.491
ModWD -SS	1.496	2.025	2.341	2.668	3.007	2.308
SS -ModWD	1.502	2.047	2.364	2.691	3.005	2.322
ModWD -WF	1.697	2.087	2.414	2.721	2.988	2.381
WF -ModWD	1.768	2.133	2.445	2.742	2.975	2.413
ModWD -STSA	1.700	2.149	2.429	2.724	2.974	2.395
STSA -ModWD	1.740	2.176	2.445	2.742	2.982	2.417
ModWD -logSTSA	1.840	2.234	2.523	2.802	3.057	2.491
logSTSA -ModWD	1.846	2.246	2.522	2.813	3.062	2.498

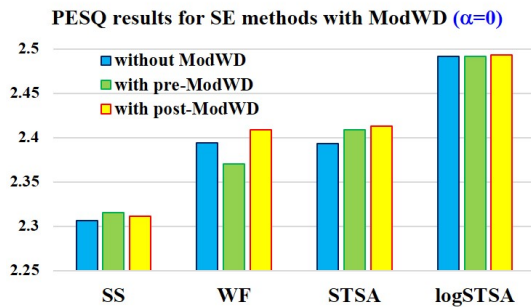


Figure 3: PESQ results averaged over five SNR cases for various SE methods with or without ModWD ( $\alpha = 0$ )

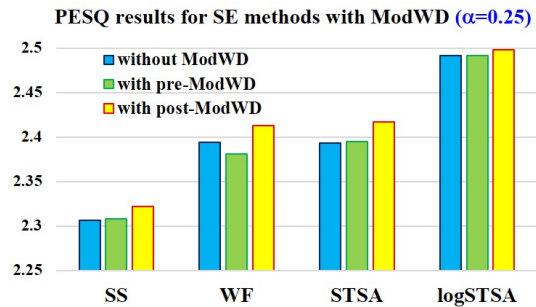


Figure 4: PESQ results averaged over five SNR cases for various SE methods with or without ModWD ( $\alpha = 0.25$ )

by ModWD is relatively insignificant because ModWD does not have an explicit noise estimation and reduction procedure as the other methods.

2. Cascading ModWD with any of the four methods gives rise to better results than the individual component method in almost all cases, revealing that ModWD is well additive to speech enhancement algorithms so as to further improve the speech quality in adverse environments.
3. For any of the four SE methods discussed here, ModWD serves as a post-processing stage better than it is used for pre-processing. A possible underlying reason for this result is ModWD tends to undermine the noise estimation accuracy of the cascaded SE method afterwards.
4. ModWD with  $\alpha = 0.25$  outperforms ModWD with  $\alpha = 0$  when they are integrated with any other SE method. However, the PESQ performance difference is marginal and less than 0.01 in most cases.

## 5. Conclusions

This study presents a DWT-based speech enhancement approach that highlights the low modulation frequency components of the spectrogram so as to reduce the noise effect. In spite of no prior knowledge of the actual distortions adopted, the presented ModWD still improves the quality of utterances in unseen noise environments and is well additive to other speech enhancement methods. As to future work, we will adopt a multi-level DWT to achieve a high resolution in modulation frequency for the analyzed spectrogram and use a validation set to learn the value of scaling factor  $\alpha$  in order to further improve the effectiveness of ModWD.

## 6. References

- [1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27(2), pp. 113120, 1979.
- [2] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International*

- Conference on Acoustics, Speech and Signal Processing*, pp. 208-211, 1979.
- [3] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2002.
- [4] C. Plapous, C. Marro, P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. on Audio, Speech and Language Processing*, 14(6), pp. 20982108, 2006.
- [5] P. Scalart, J. V. Filho, "Speech enhancement based on a pri-ori signal to noise estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 629-632, 1996.
- [6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, 32(6), pp. 11091121, 1984.
- [7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1985
- [8] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), pp. 163176, 2006.
- [9] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3), pp. 882892, 2007.
- [10] K.K. Paliwal, K.K. Wojcicki and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, Vol. 52, Issue 5, pp. 450-475, 2010.
- [11] C-C. Hsu, K-M. Cheong, J-T. Chien and T-S. Chi, "Modulation Wiener filter for improving speech intelligibility," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 370-374, 2015
- [12] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement", *Speech Communication*, Vol. 53, pp. 818-829, 2011.
- [13] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Magazine*, 1991.
- [14] S. G. Chang, B. Yu and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. on Image Processing*, vol. 9, pp. 1532-1546, Sep. 2000.
- [15] S-S. Wang, P. Lin, Y. Tsao, J-W. Hung, B. Su, "Suppression by selecting wavelets for feature compression in distributed speech recognition," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, Mar 2018
- [16] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," *European Conference on Speech Communication and Technology*, (Eurospeech), 1997.
- [17] C. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans. on Audio, Speech, and Language Processing*, pp. 257-270, 2006.
- [18] X. Xiao, E. S. Chng and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1662-1674, 2008.
- [19] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Computational Biology*, vol. 3, 2009.
- [20] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670-2680, May 1994.
- [21] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303-304, 1995.
- [22] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the 2000 Automatic Speech Recognition: Challenges for the new Millenium*, pp. 181-188, 2000
- [23] 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," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 749-752, 2001.