A PSYCHOACOUSTICALLY MOTIVATED SPEECH DISTORTION WEIGHTED MULTI-CHANNEL WIENER FILTER FOR NOISE REDUCTION

Bruno Defraene, Kim Ngo, Toon van Waterschoot, Moritz Diehl and Marc Moonen

Dept. E.E./ESAT, SCD-SISTA, Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

ABSTRACT

The aim of this paper is to improve the performance of existing speech distortion weighted multi-channel Wiener filter (SDW-MWF$_\mu$) based noise reduction (NR) algorithms. It is well known that for the SDW-MWF$_\mu$ the improved NR performance comes at the cost of higher speech distortion when a fixed speech distortion weighting factor is used. In this paper we propose two psychoacoustically motivated weighting factor selection strategies, devised to exploit masking properties of the human ear. Experimental results based on PESQ scores, SNR improvement, and signal distortion confirm that both proposed psychoacoustically motivated weighting factor selection strategies do improve the NR performance compared to using a fixed weighting factor. In some of the analyzed scenarios, the fixed weighting factor approach is even seen to degrade the PESQ scores, while the psychoacoustically motivated approaches are seen to significantly improve the PESQ scores in all of the analyzed scenarios.

Index Terms— Noise reduction, multi-channel Wiener filter, psychoacoustics, auditory masking.

1. INTRODUCTION

Additive background noise (from competing speakers, traffic etc.) is a significant problem in many speech applications, e.g. in hearing aids, hands-free mobile telephony, audio- and video-conferencing etc. Therefore both single-channel and multi-channel noise reduction (NR) algorithms have been proposed [1]. The objective of these NR algorithms is to maximally reduce the noise while minimizing speech distortion. A limitation of single-channel noise reduction is that only temporal and spectral signal characteristics can be exploited. For example, in a multiple speaker scenario (also known as the cocktail party problem) the speech (desired speaker) and the noise (competing speakers) considerably overlap in time and frequency. This makes it difficult for single-channel NR algorithms to suppress the noise without introducing speech distortion or musical noise. However, in most scenarios, the desired speaker and the noise sources are physically located at different positions. Multi-channel noise reduction algorithms can then exploit both spectral and spatial characteristics of the speech and the noise.

This research work was carried out at the ESAT Laboratory of Katholieke Universiteit Leuven, in the frame of the K.U.Leuven Research Council CoE EF/05/006 'Optimization in Engineering' (OPTEC) and PFV/10/002 (OPTEC), Concerted Research Action GOA-MaNet, the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P6/04 'Dynamical systems, control and optimization' (DYSCO) 2007-2011, Research Project IBBT, and Research Project FWO nr. G.0600.08 'Signal processing and network design for wireless acoustic sensor networks'. The scientific responsibility is assumed by its authors.

In this paper, we will focus on multi-channel NR and more specifically on so-called speech distortion weighted multi-channel Wiener filter (SDW-MWF$_\mu$) based NR [2], which provides a minimum mean square error (MMSE) estimate of the speech component in one of the input signals. The SDW-MWF$_\mu$ allows for a trade-off between noise reduction and speech distortion. A problem with the SDW-MWF$_\mu$ is related to the weighting factor (trade-off factor) which is usually fixed for each frame and for each frequency. This does not result in an optimal trade-off since speech and noise are spectrally non-stationary and in general speech contains many pauses while the noise can be continously present.

Recent work [3][4] on the SDW-MWF$_\mu$ incorporates the conditional speech presence probability (SPP) for updating the weighting factor. In speech dominant frames and frequencies it is then desirable to have less noise reduction to avoid speech distortion, while in noise dominant frames and frequencies it is desirable to have as much noise reduction as possible. This approach has shown to improve the SNR at a lower signal distortion compared to the SDW-MWF$_\mu$ using a fixed weighting factor.

In this paper we will further develop this principle of estimating a weighting factor that is updated for each frame and for each frequency by introducing a psychoacoustically motivated weighting factor, i.e. a weighting factor that is adapted based on human auditory masking properties. As such, this paper considers the inclusion of psychoacoustic principles into a multi-channel NR algorithm, as opposed to previously proposed psychoacoustically motivated single-channel NR algorithms (e.g. in [5], [6]). Experimental results with hearing aid scenarios demonstrate that the proposed SDW-MWF$_\mu$ with a psychoacoustically motivated weighting factor indeed improves the SNR, signal distortion and speech quality scores (as measured by PESQ).

The paper is organised as follows. In Section 2 the notation is introduced and the SDW-MWF$_\mu$ based NR is reviewed. The idea behind the psychoacoustically motivated weighting factor is explained in Section 3. In Section 4 experimental results are presented. The paper conclusions are given in Section 5.

2. MULTI-CHANNEL WIENER FILTER

2.1. Signal model and notation

Let $X_i(k,l), i = 1, ..., M$ denote the frequency-domain microphone signals

$$X_i(k,l) = X_i^s(k,l) + X_i^n(k,l) \tag{1}$$

where $k = 1, ..., N$ is the frequency bin index, and l is the frame index of a short-time Fourier transform (STFT), and the superscripts s and n are used to refer to the speech and the noise contribution in a

signal, respectively. Let $\mathbf{X}(k,l) \in \mathbb{C}^{M\times 1}$ be defined as the stacked vector

$$\mathbf{X}(k,l) = [X_1(k,l)\ X_2(k,l)\ ...\ X_M(k,l)]^T \quad (2)$$
$$= \mathbf{X}^s(k,l) + \mathbf{X}^n(k,l) \quad (3)$$

where the superscript T denotes the transpose. In addition, we define the speech-plus-noise, the clean speech and the noise-only correlation matrices as

$$\mathbf{R}_x(k,l) = \varepsilon\{\mathbf{X}(k,l)\mathbf{X}^H(k,l)\} \quad (4)$$
$$\mathbf{R}_s(k,l) = \varepsilon\{\mathbf{X}^s(k,l)\mathbf{X}^{s,H}(k,l)\} \quad (5)$$
$$\mathbf{R}_n(k,l) = \varepsilon\{\mathbf{X}^n(k,l)\mathbf{X}^{n,H}(k,l)\} \quad (6)$$

where $\varepsilon\{\}$ denotes the expectation operator, H denotes Hermitian transpose.

2.2. Speech distortion weighted multi-channel Wiener filter

The multi-channel Wiener filter (MWF) optimally estimates the speech signal, based on an MMSE criterion, i.e.,

$$\mathbf{W}_{\text{MWF}}(k,l) = \arg\min_{\mathbf{W}(k,l)} \varepsilon\{|X_1^s(k,l) - \mathbf{W}^H(k,l)\mathbf{X}(k,l)|^2\} \quad (7)$$

where the desired signal in this case is the (unknown) speech component $X_1^s(k,l)$ in the first microphone signal. The MWF has been extended to the SDW-MWF$_\mu$ that allows for a trade-off between noise reduction and speech distortion using a weighting factor μ [2]. If the speech and the noise signals are uncorrelated, the design criterion of the SDW-MWF$_\mu$ is given by

$$\mathbf{W}_{\text{MWF}_\mu}(k,l) = \arg\min_{\mathbf{W}(k,l)} \varepsilon\{|X_1^s(k,l) - \mathbf{W}^H(k,l)\mathbf{X}^s(k,l)|^2\} + \mu\varepsilon\{|\mathbf{W}^H(k,l)\mathbf{X}^n(k,l)|^2\} \quad (8)$$

and the SDW-MWF$_\mu$ solution is then given by

$$\boxed{\mathbf{W}_{\text{MWF}_\mu}(k,l) = \left[\mathbf{R}_s(k,l) + \mu\mathbf{R}_n(k,l)\right]^{-1}\mathbf{R}_s(k,l)\mathbf{e}_1} \quad (9)$$

where the $M \times 1$ vector $\mathbf{e}_1$ equals the first canonical vector defined as $\mathbf{e}_1 = [1\ \ 0\ \ ...\ \ 0]^T$. For μ = 1 the SDW-MWF$_\mu$ reduces to the MWF solution of (7), while for $\mu > 1$ the residual noise level will be reduced at the cost of a higher speech distortion. The output $Z(k,l)$ of the SDW-MWF$_\mu$ can then be written as

$$Z(k,l) = \mathbf{W}_{\text{MWF}_\mu}^H(k,l)\mathbf{X}(k,l). \quad (10)$$

3. INCORPORATING PSYCHOACOUSTICS

3.1. Psychoacoustical concepts

It is well-known that additive noise at certain frequencies is more perceptible than additive noise at other frequencies, and that the perceptibility is partly signal-dependent. Two phenomena of human auditory perception are responsible for this,

- The *absolute threshold of hearing* is defined as the required intensity (dB) of a pure tone such that an average listener will just hear the tone in a noiseless environment. The absolute threshold of hearing is a function of the tone frequency and has been measured experimentally [7].
- *Simultaneous masking* is a phenomenon where the presence of certain spectral energy (the masker) masks the simultaneous presence of other spectral energy (the maskee), or in other words, renders it imperceptible. In the noise reduction framework, we consider the speech frame $\mathbf{X}_1^s(l) = [X_1^s(1,l)\ \ X_1^s(2,l)\ \ ...\ \ X_1^s(N,l)]^T$ to act as the masker, and the simultaneously present noise frame $\mathbf{X}_1^n(l) = [X_1^n(1,l)\ X_1^n(2,l)\ ...\ X_1^n(N,l)]^T$ as the maskee.

Both these phenomena are taken into account in the *instantaneous masking threshold* $\mathbf{T}_1^s(l) = [T_1^s(1,l)\ T_1^s(2,l)\ ...\ T_1^s(N,l)]^T$ of the lth speech frame in the first microphone: it gives the amount of noise energy (dB) for every frequency bin k that can be masked by the speech frame. The instantaneous masking threshold basically tells us that in order to render the residual noise inaudible in the presence of the speech, we need to make its level equal or lower than the speech masking threshold $T_1^s(k,l)$. By making the weighting factor μ in the SDW-MWF$_\mu$ formulation (8) time and frequency dependent, i.e. $\mu(k,l)$, and furthermore dependent on the masking treshold $T_1^s(k,l)$, it is now possible to judiciously trade-off residual noise and speech distortion from a perceptual point of view.

3.2. Psychoacoustical speech distortion weighting factor

Intuitively, it is clear that a higher masking threshold $T_1^s(k,l)$ should result in a lower weighting factor $\mu(k,l)$ and vice versa:

- When $T_1^s(k,l)$ is low, more emphasis should be put on noise reduction (high $\mu(k,l)$) because of the low noise masking capabilities of the speech frame in this frequency bin. This comes at the cost of a higher speech distortion.
- When $T_1^s(k,l)$ is high, less emphasis should be put on noise reduction (low $\mu(k,l)$) because of the high noise masking capabilities of the speech frame in this frequency bin. This allows to keep the speech distortion low, which is perceptually beneficial as we note that frequency regions where $T_1^s(k,l)$ is high typically coincide with regions of speech presence.
- When $T_1^s(k,l)$ exceeds the noise level $X_1^n(k,l)$, no noise reduction should be performed ($\mu(k,l) = 0$), as the noise is already masked by the speech.

Based on the considerations above, we now propose two different weighting factor selection strategies.

Selection strategy 1:

A first selection strategy is purely based on $T_1^s(k,l)$, i.e.

$$\boxed{\mu_{p_1}(k,l) = \begin{cases} \alpha\, e^{\beta T_1^s(k,l)}, & T_1^s(k,l) \le \nu \\ 0, & T_1^s(k,l) > \nu \end{cases}} \quad (11)$$

with parameters (α,β,ν). As $\mu_{p_1}(k,l)$ should be positive and monotonously decreasing for increasing $T_1^s(k,l)$, α is necessarily positive and β is necessarily negative. The parameter ν can be chosen as an a priori estimate of the average noise level.

Selection strategy 2:

If additionally, the noise $X_1^n(k,l)$ is assumed known or a good estimate thereof is available, we propose the following selection strategy, now mapping the noise-to-mask-ratio $NMR(k,l) = 20\log|X_1^n(k,l)| - T_1^s(k,l)$ to $\mu_{p_2}(k,l)$,

$$\mu_{p_2}(k,l) = \begin{cases} \gamma NMR(k,l)^{\delta} + \epsilon, & NMR(k,l) \geq 0 \\ 0, & NMR(k,l) < 0 \end{cases} \quad (12)$$

with parameters (γ,δ,ϵ). As $\mu_{p_2}(k,l)$ should be positive and monotonously increasing for increasing $NMR(k,l)$, δ is necessarily positive. As opposed to the first selection strategy, this selection strategy will guarantee that no noise reduction will be performed ($\mu_{p_2}(k,l) = 0$) whenever the noise is already masked by the speech ($NMR(k,l) < 0$).

3.3. Instantaneous masking threshold calculation

The instantaneous masking threshold is calculated using part of the ISO/IEC 11172-3 MPEG-1 Layer 1 psychoacoustic model 1. A complete description of the operation of this psychoacoustic model is beyond the scope of this paper (we refer the reader to [7]). We will outline the relevant steps in the computation of the instantaneous masking threshold $\mathbf{T}_1^s(l)$:

1. *Identification of tonal and non-tonal maskers*: It is known from psychoacoustic research that the tonality of a masking component has an influence on its masking properties. For this reason it is important to discriminate between tonal and non-tonal maskers in the spectrum $\mathbf{X}_1^s(l)$. In a first phase, tonal maskers are identified at local maxima of the PSD: energy from three adjacent spectral components centered at the local maximum is combined to form a single tonal masker. In a second phase, a single non-tonal masker per critical band is formed by addition of all the energy from the spectral components within the critical band that have not contributed to a tonal masker.
2. *Decimation of maskers*: In this step, the number of maskers is reduced using two criteria. First, any tonal or non-tonal masker below the absolute threshold of hearing is discarded. Next, any pair of maskers occurring within a distance of 0.5 Bark is replaced by the stronger of the two.
3. *Calculation of individual masking thresholds*: an individual masking threshold is calculated for each masker in the decimated set of tonal and non-tonal maskers, using fixed psychoacoustic rules. Essentially, the individual masking threshold depends on the frequency, loudness level and tonality of the masker.
4. *Calculation of global masking threshold*: Finally, the global masking threshold $\mathbf{T}_1^s(l)$ is calculated by a power-additive combination of the tonal and non-tonal individual masking thresholds, and the absolute threshold of hearing.

To explore the full potential of using masking thresholds, in this paper we make the assumption that the speech masking threshold $\mathbf{T}_1^s(l)$ can be estimated based on the speech components in the first microphone signal, $\mathbf{X}_1^s(l)$. In practical implementations, the masking threshold will of course have to be estimated based on the noisy microphone signals. Different strategies for estimating the masking threshold based on the noisy speech signals can be envisaged: in the context of psychoacoustically motivated single-channel NR, it was proposed to first compute a rough estimate of the clean speech signal with a simple power spectral subtraction scheme, after which the masking threshold is calculated [5]. Alternatively, one could use the estimate of the clean speech correlation matrix $\mathbf{R}_s(k,l)$ to extract the clean speech PSD of the first microphone signal, and calculate the masking threshold based on this PSD estimate.

4. EXPERIMENTAL RESULTS

4.1. Experimental set-up

Simulations have been performed with a 2-microphone (with an intermicrophone distance of approximately 1cm) behind-the-ear hearing aid mounted on a CORTEX MK2 manikin such that the head-shadow effect is included. The loudspeakers (FOSTEX 6301B) are positioned at 1 meter from the center of the head. The reverberation time $T_{60} = 0.61s$. The speech signal consists of male sentences from the Hearing in Noise Test (HINT) database for the measurement of speech reception thresholds in quiet and in noise, and the noise signals consist of a multi-talker babble from Auditory Tests (Revised), Compact Disc, Auditec. The signals are sampled at 16kHz. An FFT length of 128 is used with 50% overlap and Hanning windowing. Two different input SNRs are considered, namely -5dB and 0dB. Four spatial scenarios are considered, where the spatial angle of the single noise source is set to 30°,60°,90° and 120°, with the speech source at 0°. Five different weighting factor selection strategies are considered for comparative evaluation:

- Fixed $\mu = 1$, $\mu = 3$, $\mu = 5$.
- Psychoacoustically motivated $\mu_{p_1}(k,l)$ as defined in (11), with (α,β,ν)=$(4.374, -0.0282, 40)$.
- Psychoacoustically motivated $\mu_{p_2}(k,l)$ as defined in (12), with (γ,δ,ϵ)=$(0.1226, 0.8598, 0.9405)$.

4.2. Performance measures

To assess the noise reduction performance the intelligibility-weighted SNR [8] is used which is defined as

$$\Delta\text{SNR}_{\text{intellig}} = \sum_i I_i(\text{SNR}_{i,out} - \text{SNR}_{i,in}) \quad (13)$$

where I_i is the band importance function defined in ANSI S3.5-1997 [9] and where $\text{SNR}_{i,out}$ and $\text{SNR}_{i,in}$ represent the output SNR and the input SNR (in dB) of the *i*-th band, respectively.

For measuring the signal distortion a frequency-weighted log-spectral signal distortion (SD) is used defined as

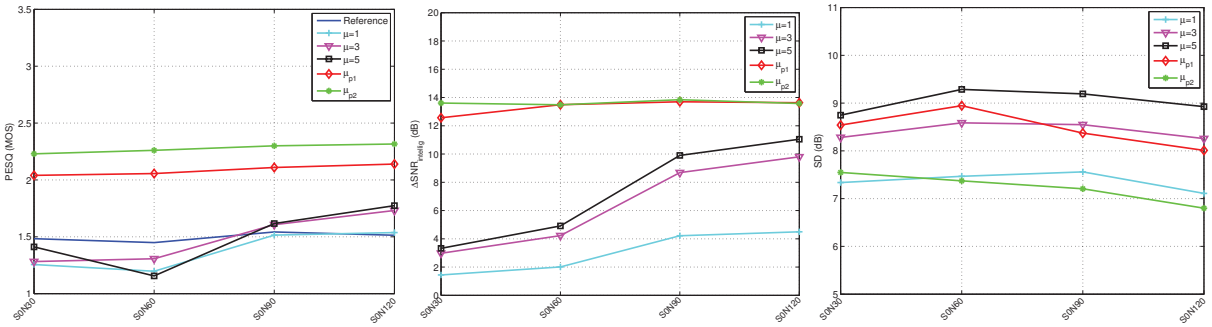
$$\text{SD} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\int_{f_l}^{f_u} w_{ERB}(f)\Big(10log_{10}\frac{P_{out,k}^s(f)}{P_{in,k}^s(f)}\Big)^2 df} \quad (14)$$

where K is the number of frames, $P_{out,k}^s(f)$ is the output power spectrum of the kth frame, $P_{in,k}^s(f)$ is the input power spectrum of the kth frame and f is the frequency index. The SD measure is calculated with a frequency-weighting factor $w_{ERB}(f)$ giving equal weight for each auditory critical band, as defined by the equivalent rectangular bandwidth (ERB) of the auditory filter.

To evaluate the perceptual quality of the processed speech, PESQ [10] is used. The PESQ algorithm is presented with the clean, unprocessed reference microphone speech signal and the processed noisy signal, and calculates a Mean Opinion Score (MOS) on a scale from 1 to 5, thus predicting the subjective speech quality of the processed signal.

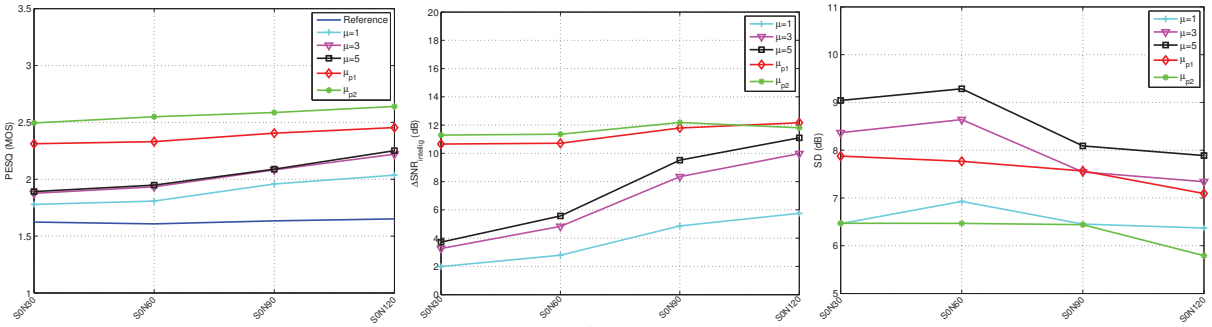
4.3. Results and discussion

In Fig. 1 and Fig. 2, simulation results for the SDW-MWF$_\mu$ with different weighting factor selection strategies are shown for scenarios with an input SNR of -5dB and 0dB, respectively. In these figures, S0Nx denotes a spatial scenario with the speech source at 0°, and



(a) PESQ (b) SNR improvement (c) SD

Fig. 1. Input SNR=-5 dB



(a) PESQ (b) SNR improvement (c) SD

Fig. 2. Input SNR=0 dB

the single noise source at x°. A first observation is that the proposed psychoacoustically motivated weighting factor selection strategies μ_{p1} and μ_{p2} result in significantly higher PESQ scores and SNR improvement, as compared to fixed μ strategies. Moreover, the higher SNR improvement does not come at the cost of a higher speech distortion, which is comparable to and often even lower than using fixed μ strategies. This observation is seen to hold for both considered input SNRs, and for all considered spatial scenarios. A second observation is that for the spatial scenarios S0N30 and S0N60 with an input SNR of -5dB, the PESQ score is even degraded by using a fixed μ compared to the reference PESQ score (solid line), while μ_{p1} and μ_{p2} significantly improve the PESQ scores in all of the analyzed scenarios. A third observation is that in general μ_{p2} is seen to slightly outperform μ_{p1} for all three performance measures.

5. CONCLUSION

In this paper we have proposed two psychoacoustically motivated weighting factor selection strategies for the SDW-MWF$_{\mu}$, and investigated their comparative performance to fixed weighting factor strategies. Experimental results with hearing aid scenarios demonstrate that both proposed psychoacoustically motivated SDW-MWF$_{\mu}$ approaches significantly outperform fixed weighting factor strategies in terms of the objective measures PESQ, SNR improvement, and signal distortion. Moreover, for some scenarios, the fixed weighting factor approaches are seen to degrade the PESQ scores, while the psychoacoustically motivated approaches are seen to significantly improve the PESQ scores for all of the analyzed scenarios.

6. REFERENCES

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

[2] A. Spriet, M. Moonen, and J. Wouters, "Stochastic gradient based implementation of spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction in hearing aids," *IEEE Trans. on Sig. Proc.*, vol. 53, no. 3, pp. 911–625, Mar. 2005.

[3] K. Ngo, A. Spriet, M. Moonen, J. Wouters, and S. H. Jensen, "Incorporating the conditional speech presence probability in multi-channel Wiener filter based noise reduction in hearing aids," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 930625, 11 pages, 2009, doi:10.1155/2009/930625.

[4] K. Ngo, M. Moonen, J. Wouters, and S. H. Jensen, "A flexible speech distortion weighted multi-channel Wiener filter for noise reduction in hearing aids," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2011.

[5] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126 – 137, Mar. 1999.

[6] Y.Hu and P.C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 457–465, Sept. 2003.

[7] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, Apr. 2000.

[8] J. E. Greenberg, P. M. Peterson, and P. M. Zurek, "Intelligibility-weighted measures of speech-to-interference ratio and speech system performance," *Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 3009–3010, Nov. 1993.

[9] Acoustical Society of America, "ANSI S3.5-1997 American National Standard Methods for calculation of the speech intelligibility index," June 1997.

[10] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), a new method for speech quality assessment of telephone networks and codecs," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, pp. 749–752, 2001.