Noise-Adaptive Perceptual Weighting in the AMR-WB Encoder for Increased Speech Loudness in Adverse Far-End Noise Conditions

Emma Jokinen
Department of Signal Processing and Acoustics
Aalto University
Espoo, Finland
emma.jokinen@aalto.fi

Tom Bäckström
International Audio Laboratories Erlangen
Friedrich-Alexander University (FAU)
Germany
tom.backstrom@audiolabs-erlangen.de

***Abstract*—In mobile communications, environmental noise often reduces the quality and intelligibility of speech. Problems caused by far-end noise, in the sending side of the communication channel, can be alleviated by using a noise reducing pre-processing stage before the encoder. In this study, a modification increasing the robustness of the encoder itself to background noise is proposed. Specifically, by using information already present in the encoder, the proposed method adjusts the perceptual weighting filter based on the characteristics of the noise to increase the prominence of the speech over the background noise. To evaluate the performance of the enhancement, the modification is implemented in the adaptive multi-rate wideband encoder and compared to the standard AMR-WB encoder in subjective tests. The results suggest that while the proposed modification increases the loudness of speech without affecting the quality significantly, for some female speakers the standard encoder is preferred over the enhancement.**

***Index Terms*—Speech coding, AMR-WB, far-end noise, perceptual weighting, loudness**

I. Introduction

Adverse background noise conditions are common in mobile communications. Environmental noise can be present in either side of the communication channel and both the quality and intelligibility of the communication can be affected negatively. Several types of pre- and post-processing techniques suitable for mobile devices have been previously proposed to alleviate the problem. For instance, in the speaker's end the effects of noise can be diminished by utilizing noise suppression (e.g., [1], [2], [3], [4], [5]) as a pre-processing step. Additionally, in the receiving device, the intelligibility of the speech can be increased over near-end noise in the listener's surroundings with the utilization of post-processing techniques (e.g., [6], [7], [8], [9], [10], [11], [12]). However, whether additional enhancement techniques are used, depends highly on the phone manufacturer. Therefore, the performance of the speech codec alone in the presence of degradations is very important. Furthermore, if these enhancement techniques are integrated into the operations of the encoder or decoder itself, the overall computational cost and delay in the mobile device might be decreased.

The focus of this paper is on the far-end noise scenario, where the speech signal is corrupted by noise before encoding. Since the codec now has to encode both the desired speech signal and the undesired distortions, the coding problem is more complicated because the signal now consists of two sources, and that will decrease encoding quality. Even if the combination of the two sources could be encoded with the same quality as a clean speech signal, the speech part would still have lower quality than the clean signal.The lost encoding quality is not only perceptually annoying, but importantly, it also increases listening effort and in the worst case, decreases the intelligibility of the decoded signal.

For this problem setting, the conventional approach for noise suppression is to apply a separate pre-processing block with the purpose of removing noise before coding. However, by separating the noise suppression and the encoding to separate blocks, two main disadvantages arise. First, since the noise suppressor will generally not only remove noise but also distort the desired signal, the codec will thus attempt to encode a distorted signal accurately. The codec will therefore have a wrong target and both efficiency and accuracy are lost. This can also be seen as a case of tandeming problem, where subsequent blocks produce independent errors which add up. This problem was addressed in [13], where an optimization of the pre-processing noise reduction stage based on the impact on the encoder performance was proposed. Similarly, in [14], the tandeming of different noise reduction and coding techniques was evaluated and suggestions on optimal combinations were made. However, in both cases the noise reduction was still considered to be a pre-processing step instead of being integrated fully into the encoder which results in a higher computational cost and delay than in an embedded solution. Additionally, by joint noise suppression and encoding, such tandeming problems can be completely avoided. A partially integrated coding/enhancement scheme with low-delay was studied in [15]. Approaches where the enhancement is fully integrated in to the encoder have been proposed, for instance, in [16] and in [17], where noise reduction is done in compressed domain by optimally modifying the fixed and adaptive

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

codebook gains. Although both of the suggested methods were mostly intended for noise reduction in the network, in principle these kind of compressed domain techniques could be embedded into the encoder as well.

In this study, a noise-adaptive modification to the encoder designed to reduce the degradation caused by far-end noise is proposed. The main idea is to adjust the perceptual weighting filter based on the characteristics of the noise. In other words, the far-end noise is not explicitly suppressed but the perceptual objective function is changed such that the accuracy is higher in parts where the signal-to-noise ratio (SNR) is the best. Equivalently, the purpose is to decrease signal distortion at those parts where SNR is high. Those parts of the signal which have low SNR are thus transmitted with less accuracy but since they contain mostly noise, encoding them accurately is not considered important. Importantly, the proposed method does not in general provide the most accurate possible representation of the input signal, but the target is to transmit such parts of the speech signal that increase its prominence over the background noise. Specifically, the timbre of the signal might be changed, but in such a way that the transmitted speech signal sounds louder and is, thus, better in terms of intelligibility and listening effort than the accurately transmitted signal.

The proposed method uses information already computed in the encoder as a part of the standard functionality and therefore, the additional computational load is small. The introduced modification is implemented in the adaptive multi-rate wideband (AMR-WB) encoder [18] and evaluated in comparison to the standard AMR-WB encoder with subjective pair comparison tests using two SNR levels of additive, far-end background noise.

II. Proposed method

Most speech codecs, including the AMR-WB codec, use algebraic code-excited linear prediction (ACELP) for parametrizing the speech signal. This means that first the contribution of the vocal tract, $A(z)$, is estimated with linear prediction and removed and after this, the residual signal is parametrized using an algebraic codebook. For finding the best codebook entry, a perceptual distance between the original residual and the codebook entries is minimized. The perceptual distance function can be written as

$$||\boldsymbol{W}\boldsymbol{H}(x - \hat{x})||^2, \tag{1}$$

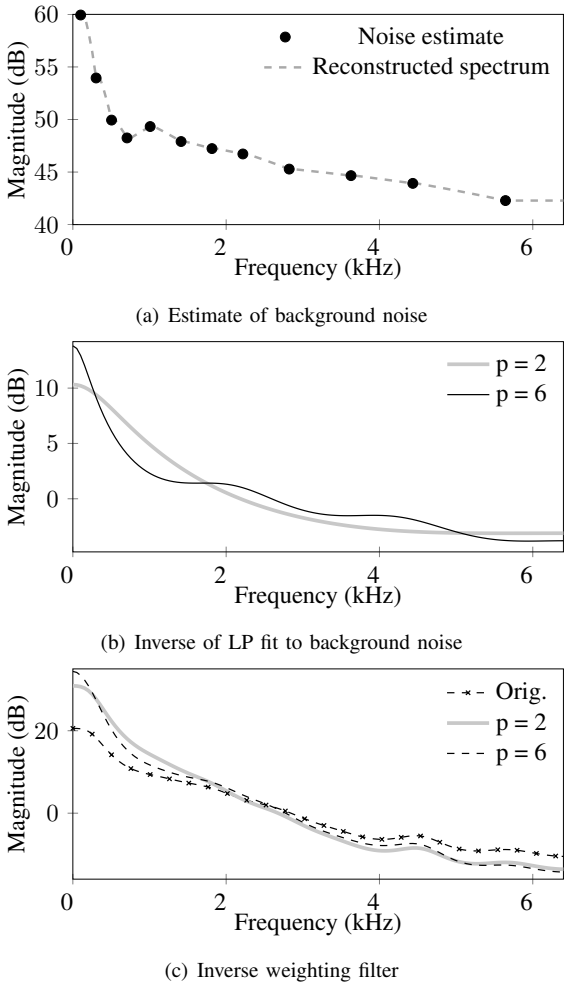
where x and $\hat{x}$ are the original and quantized residuals, $\boldsymbol{W}$ and $\boldsymbol{H}$ are the convolution matrices corresponding, respectively, to $H(z) = 1/\hat{A}(z)$, the quantized vocal tract synthesis filter and $W(z)$, the perceptual weighting, which is typically chosen as

$$W(z) = A(z/\gamma_1)H_{\mathrm{de-emph}}(z) \tag{2}$$

with $\gamma_1 = 0.92$. Furthermore,

$$H_{\mathrm{de-emph}}(z) = 1/(1 - \beta_1 z^{-1}) \tag{3}$$

with $\beta_1 = 0.68$ is the de-emphasis filter which is used to compensate for the pre-emphasis done in the beginning of the encoding. The residual x has been computed with the quantized vocal tract analysis filter.



(a) Estimate of background noise

(b) Inverse of LP fit to background noise

(c) Inverse weighting filter

Fig. 1. An example of the construction of the noise-adaptive weighting filter for car noise with average SNR of -5 dB. In (a), the original background estimate computed by the encoder and the reconstructed spectrum are shown. Figure (b) depicts the spectra of the inverses of the linear prediction fits ($1/A_{\mathrm{BCK}}(z)$) for the background noise estimate with different prediction orders. Finally, (c) displays the frequency responses of the inverses of the original and the proposed weighting filters with different prediction orders.

In the application scenario of this study, the incoming speech signal contains additive far-end noise. Thus, the signal is

$$y(t) = s(t) + n(t). \tag{4}$$

In this case, both the vocal tract model, $A(z)$, and the original residual contain noise. For this study, the noise in the vocal tract model is ignored and the focus is placed on the noise in the residual. The idea behind the proposed modification is to guide the perceptual weighting such that the effects of the additive noise are reduced in the quantization of the residual. Whereas normally the target is to make the error between the original and the quantized residual to resemble the speech spectral envelope, in this case the aim is to minimize the error in the region which is considered more robust to noise. In other words, the frequency components that are less

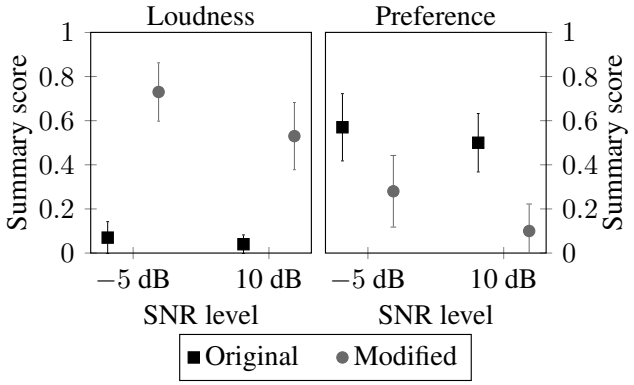


Fig. 2. The means of the summary scores for both loudness and listening preference as well as their 95% confidence intervals for both SNR conditions. The scores have been averaged over all the speakers in the test. The methods being compared in the test were the original encoder and the modified encoder.

corrupted by the noise should be quantized with less error whereas components with low magnitudes which are likely to contain errors from the noise should have a lower weight in the quantization process.

To take into account the effect of noise on the desired signal, an estimate of the noise signal is needed first. Noise estimation is a classic topic for which many methods exist. Here, a low-complexity method, which uses information already existing in the encoder, is utilized. An estimate of the shape of the background noise is stored for the voice activity detection (VAD). This estimate contains the level of the background noise in 12 frequency bands with increasing width. A spectrum is constructed from this estimate by mapping it to a linear frequency scale by using interpolation between the original data points. An example of the original background estimate and the reconstructed spectrum is shown in Fig. 1. From the reconstructed spectrum, the autocorrelation is computed and used to derive the pth order linear prediction (LP) coefficients with the Levinson-Durbin recursion. Examples of the obtained LP fits with $p = 2$ and $p = 6$ are shown in Fig. 1. As seen from the figure, low prediction order captures a rough spectral envelope of the noise while the model with order $p = 6$, already contains some finer details.

The obtained LP fit, $A_{\mathrm{BCK}}(z)$ can be used as a part of the weighting filter in the computation of the best codebook entry. Finally, the new weighting filter will be

$$W(z) = A(z/\gamma_1)A_{\mathrm{BCK}}(z/\alpha)H_{\mathrm{de-emph}}(z). \qquad (5)$$

Parameter α can be used to adjust the amount of noise-dependent weighting. If $\alpha \to 0$, the effect is small, while for $\alpha \approx 1$ a high level of noise-dependent weighting is obtained.

In Fig. 1, an example of the inverse of the original weighting filter as well as the inverses of the proposed weighting filters with $\alpha = 1$ and prediction orders $p = 2$ and $p = 6$ is shown. For the figure, the de-emphasis filter, $H_{\mathrm{de-emph}}(z)$, has not been used. While the difference between the original weighting filter and the modified weighting filters is quite large, the differences between the two modified filters with prediction orders $p = 2$ and $p = 6$ is relatively small. Furthermore, because the background noise estimate computed in the encoder contains few data points, using an LP model that captures details of the reconstructed noise spectrum is not necessary but simply increases the computational load of the proposed modification. For these reasons, the prediction order in the proposed method is set to $p = 2$. Additionally, for the evaluations done in this study, parameter α was set to 1 which means that the weighting filter is always adapted fully to the background noise conditions.

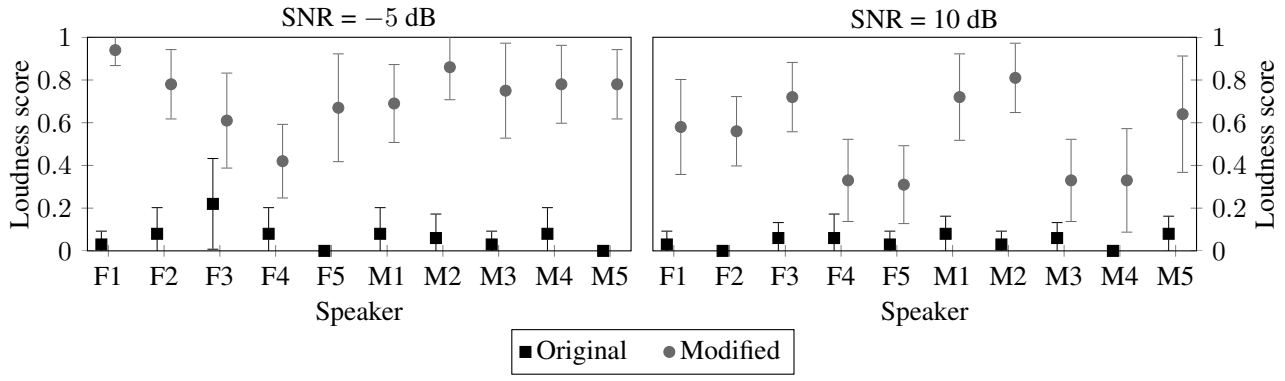
III. Subjective evaluation

A subjective listening test was organized to evaluate the performance of the modified encoder in comparison to the original encoder. The test consisted of a pair comparison test with two questions regarding the subjective loudness and listening preference of the samples in noisy conditions. Loudness was selected as an attribute in the test instead of intelligibility or listening effort because listeners can have difficulties judging intelligibility or listening effort in a pair comparison test. This is especially true in background noise conditions where the intelligibility approaches 100%.

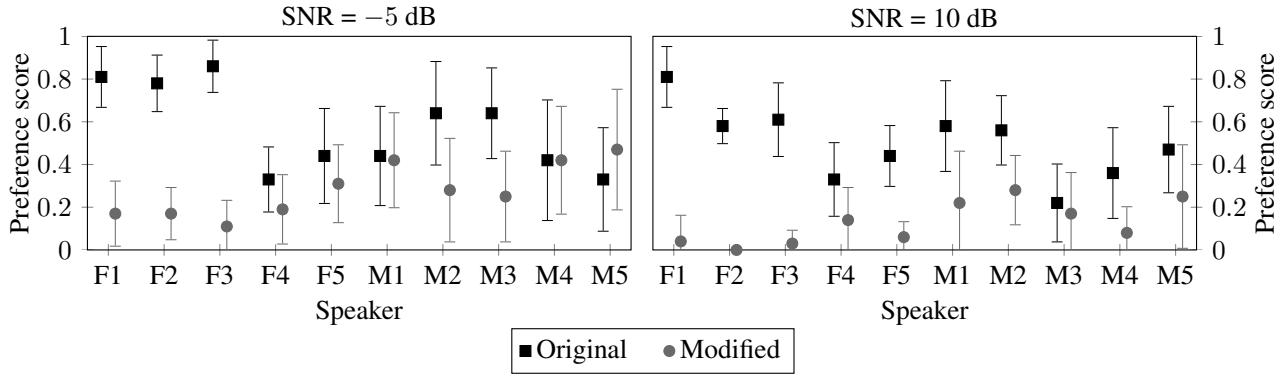
The background noise refers here to a far-end noise condition which means that the degrading environmental noise is on the sending side of the communication channel. Thus, the encoding and decoding are both affected by the noise. The SNR levels for the test were selected such that in addition to the degradation in quality, the intelligibility would also be negatively affected. Typically in quality evaluations of coding standards, the SNR levels for far-end background noise are around 15 to 20 dB (e.g., [19], [20]) which does not affect intelligibility or listening effort adversely. Therefore, in this test car noise with two SNR levels, 10 dB and -5 dB, was used. The first SNR was selected from a typical operating range which affects mostly quality and the second SNR level was chosen to be much lower in order to create noise conditions where the listening effort would be increased.

The speech material in the test consisted of Finnish sentence material from five male and five female speakers. The sentences contained each three words and had an average duration of approximately 2 seconds. All speech samples were first preprocessed to correspond to wideband telephone speech by first filtering at 48-kHz rate with the HP50 filter, which is a high-pass filter simulating mobile device input characteristics [21]. Then the samples were downsampled to 16 kHz and level adjusted to -26 dBov with SV56 [22]. After this, stationary car noise was added to the samples according to the SNR level under evaluation and the resulting noisy speech signal was encoded with either the original or the modified AMR-WB encoder with a rate of 23.05 kpbs. Finally, the encoded signal was decoded using the standard AMR-WB decoder.

Eleven normal-hearing listeners, all native speakers of Finnish, participated in the listening tests. The age of the listeners ranged from 26 to 47 with an average of 31 years. The tests took place in a sound-proofed listening booth using



(a) Loudness



(b) Listening preference

Fig. 3. The means of the summary scores for (a) loudness and (b) listening preference and their 95% confidence intervals for all the speakers for both SNR conditions. The speakers from F1 to F5 are female and from M1 to M5 male. The methods being compared in the test were the original encoder and the modified encoder.

Sennheiser HD 650 headphones. The test was divided into two parts where the first part contained samples with -5 dB SNR level and the second part samples with 10 dB SNR level. In the beginning of the test session, a short practice test was given to the participants. The A-weighted sound pressure level was set to 65 dB and kept constant throughout the tests.

In the pair comparison test, the listeners were able to freely listen to two samples, A and B, and were asked to answer two questions:

Q1: *"Which sample sounds louder?"*
Q2: *"Which sample do you prefer to listen to?"*

They were asked to choose one of the options: *A*, *B* or *No difference* and instructed to select *No difference* if they had no preference even if they heard a difference between the samples. All pairs of samples were presented in both orders and additionally, null pairs, where both the samples were the same, were used to control the quality and consistency of the listeners.

A. Results

Before the analysis, the listeners were checked for consistency in the pair comparison test using the scores of the null pairs. If over half of the null pairs in the tests were rated "A" or "B" instead of "No difference", the listener was discarded. Based on this quality control, two listeners out of 11 were removed from further analysis. After this, the results of the pair comparison test were analyzed separately for loudness and for listening preference. The summary scores for each were first evaluated by computing the number of times each method was selected in all comparisons it was involved in for all the speakers in the test. These scores were then analyzed with a three-way analysis of variance with the method (original, modified), speaker (female speakers F1-F5, male speakers M1-M5) and SNR level (-5 dB, 10 dB) as fixed factors.

The analysis of the summary scores on loudness showed that the method $[(F(1,9) = 252.9, p < 0.01)]$, the SNR level $[(F(1,9) = 1.53, p < 0.05)]$ as well as the interaction between the method and the SNR level $[(F(1,9) = 54.6, p < 0.05)]$ had a significant effect. The post-hoc tests using Tukey's method indicated that the modified encoder received overall higher ratings than the original encoder. Although the same ranking was observed in both SNR levels, the difference between the loudness ratings of the two encoders was larger in the lower SNR condition.

The summary scores on listening preference were affected by the method $[(F(1,9) = 81, p < 0.01)]$, the SNR level $[(F(1,9) = 9, p < 0.05)]$ as well as the interaction between the method and the speaker $[(F(9,9) = 4.06, p < 0.05)]$. The post-hoc tests revealed that while the original encoder was rated overall higher than the modified encoder, this difference

was only significant with the female speakers F1-F3. For the other speakers, no significant differences were found between the original and modified encoders in terms of listening preference. The results both in terms of loudness and listening preference are visualized in Figs. 2 and 3.

IV. Conclusion

An enhancement of the perceptual weighting filter of the AMR-WB encoder in the presence of far-end background noise was introduced. The proposed technique aims to increase the robustness of the encoding in noise by taking advantage of information already present in the encoder, thus adding a relatively small computational load to the encoding. The goal of the proposed enhancement is not to explicitly suppress the far-end noise present in the signal, but to encode the signal such that the prominence of the speech is increased over the background noise. The performance of the enhancement in comparison to the original encoder was evaluated in subjective tests in terms of loudness and listening preference with two levels of far-end background noise.

The results suggest that the proposed modification is able to increase the loudness of speech without affecting the quality significantly for most speakers. However, for some female speakers the standard encoder received significantly higher listening preference ratings which suggests that there are individual differences on how the enhancement works. Based on informal listening, the enhanced speech sounds in some cases overly sharp which reduces the quality and listening comfort. This could be adjusted by controlling the effect that the noise-adaptive filter has on the perceptual weighting. In the evaluations done in this study, the perceptual weighting was always adapted fully according to the background noise.

Individual differences between how the enhancement works can also be related to the functioning of the VAD in the encoder. Depending on the speaking style of the individual, the background noise estimate, which is used for the proposed enhancement, is updated differently. For some speakers the background noise estimate is rarely updated and is thus not very efficient in adapting the perceptual weighting to the noise. The reliability of the VAD decisions also decreases for all speakers as the noise level increases. In most cases this does not seem to be a problem but further work is required in order to resolve where the individual differences arise from.

In conclusion, the proposed method shows promising results in terms of loudness increase in difficult noise conditions with a minimal increase in computational cost and delay. Furthermore, the method is conveniently applicable to any codec employing a perceptual model and further work will therefore also include evaluation of the method using the recently standardized Enhanced Voice Services (EVS) codec.

Acknowledgment

The International Audio Laboratories Erlangen (AudioLabs) is a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

References

[1] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*, Springer-Verlag, Heidelberg, 2005.

[2] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 1693–1696.

[3] Z. Koldovsky, P. Tichavsky, and D. Botka, "Noise reduction in dual-microphone mobile phones using a bank of pre-measured target-cancellation filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 679–683.

[4] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 12, pp. 1235–1239, 2013.

[5] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[6] B. Sauert, G. Enzner, and P. Vary, "Near end listening enhancement with strict loudspeaker output power constraining," in *Proc. IWAENC*, 2006.

[7] J.L. Hall and J.L. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *J. Acoust. Soc. Amer.*, vol. 127, no. 1, pp. 280–285, 2010.

[8] T.-C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, 2012.

[9] E. Jokinen, P. Alku, and M. Vainio, "Lombard-motivated post-filtering method for the intelligibility enhancement of telephone speech," in *Proc. Interspeech*, 2012.

[10] C.H. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, 2013.

[11] Y. Tang and M. Cooke, "Energy reallocation strategies for speech enhancement in known noise conditions," in *Proc. Interspeech*, 2010, pp. 1636–1639.

[12] H. Schepker, J. Rennies, and S. Doclo, "Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression," in *Proc. Interspeech*, 2013, pp. 3577–3581.

[13] R. Martin, I. Wittke, and P. Jax, "Optimized estimation of spectral parameters for the coding of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2000, pp. 1479–1482.

[14] D. Virette, P. Scalart, and C. Lamblin, "Analysis of background noise reduction techniques for robust speech coding," in *Proc. Eusipco*, 2002, pp. 1–4.

[15] R. Martin, H.-G. Kang, and R.V. Cox, "Low delay analysis/synthesis schemes for joint speech enhancement and low bit rate speech coding.," in *Proc. EUROSPEECH*, 1999.

[16] H. Taddei, C. Beaugeant, and M. de Meuleneire, "Noise reduction on speech codec parameters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2004, pp. 497–500.

[17] R. Chandran and D.J. Marchok, "Compressed domain noise reduction and echo suppression for network speech enhancement," in *Proc. IEEE Midwest Symp. Circ. Syst.*, 2000, pp. 10–13.

[18] 3rd Generation Partnership Project, Valbonne, France, *Specification TS 26.171: Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description*, 2012, version 11.0.0.

[19] 3rd Generation Partnership Project, Valbonne, France, *Specification TR 26.952: Codec for Enhanced Voice Services (EVS); Performance characterization*, 2015, version 12.2.0.

[20] A. Rämö and H. Toukomaa, "Subjective quality evaluation of the 3GPP EVS codec," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2015, pp. 5157–5161.

[21] Int. Telecommun. Union, Geneva, Switzerland, *Recommendation G.191: Software tools for speech and audio coding standardization*, September 2005.

[22] Int. Telecommun. Union, Geneva, Switzerland, *Recommendation P.56: Objective measurement of active speech level*, March 1993.