# Frequency-adaptive post-filtering for intelligibility enhancement of narrowband telephone speech

*Emma Jokinen, Marko Takanen, Paavo Alku*

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

## Abstract

Post-filtering methods are used in mobile communications to improve the quality and intelligibility of speech. This paper introduces a frequency-adaptive post-filtering algorithm that selects from a predefined set of filters the one that reallocates the largest amount of speech energy from low to high frequencies. The proposed method and another post-filtering technique were compared to unprocessed speech in subjective listening tests in terms of intelligibility. The results indicate that the proposed method outperforms the reference method in difficult noise conditions.

**Index Terms**: Speech enhancement, telephone speech, post-filtering, intelligibility

## 1. Introduction

Post-processing is used in mobile communications to enhance the quality and intelligibility of speech that has been degraded by quantization noise or acoustical background noise in the receiving side of the communication channel. *Post-filtering*, a typical form of post-processing, refers to filtering the decoded speech signal at the mobile phone receiver with an adaptive filter. A conventional method aimed at improving the perceptual quality, introduced in [1], emphasizes spectral peaks and attenuates noise components at spectral valleys. This approach has been developed further in several studies, for instance, a modified version with a better adaptability to varying noise conditions was proposed in [2]. Many previous studies on post-filtering focus on improving the quality of speech and they are typically tested in low to moderate noise conditions [1, 2, 3]. However, mobile phones are also used in situations with difficult background noise conditions. In these circumstances, quality is no longer essential and speech intelligibility becomes the most important factor affecting speech communication.

Post-processing methods have also been developed to improve speech intelligibility. Some of these algorithms are based on the optimization of different objective measures which are known to correlate with subjective intelligibility. For instance, in [4, 5], the maximization of the speech intelligibility index (SII) was utilized to find the optimal gains for the sub-bands of the unprocessed speech signal. No subjective tests were, however, conducted. In [6], selective boosting of frequency regions with low signal-to-noise ratios (SNRs) was used to reallocate speech energy. This work was extended in [7] towards a more robust implementation with optimization schemes for different noise types with the glimpse proportion (GP) measure. The optimization was done offline and therefore, the method is not suitable for real-time implementation. In [8], the maximization of the likelihood of noisy speech given a statistical model of clean speech was utilized to derive the optimal band-energy gains. This technique, however, called for a transcription of the noisy speech which makes the algorithm applicable for a restricted domain only.

Another approach to enhance intelligibility is to take advantage of the phenomena that are used by talkers naturally when they are trying to overcome communication barriers, such as noise. For instance in [9], energy reallocation was utilized to transfer energy from voiced sounds to unvoiced utterances. In [10], adaptive spectral shaping, aimed at sharpening the formants and reducing the spectral tilt, was combined with dynamic range compression. The approach was shown to improve both objective and subjective intelligibility. In [11], spectral tilt reduction and formant sharpening were combined in a post-filtering method that was shown to improve intelligibility in various noise conditions. Energy reallocation can also be used in post-filtering algorithms ([12, 13, 14]) by utilizing a high-pass type post-filter to attenuate the low frequency regions and to enhance higher frequencies, effectively transferring more energy to the high frequency region. While this post-filtering technique has been shown to result in improved intelligibility, it might be difficult to adapt to speaker dependent voice characteristics. In addition, strong high-pass type filtering can reduce the naturalness of speech [15] and change the voice quality of the speaker [13]. Therefore, it would be beneficial to have a similar modification maintaining the naturalness that would still provide the same improvement in terms of intelligibility.

This study introduces a frequency-adaptive post-filtering method aimed at improving the intelligibility of narrowband telephone speech. Although the usage of wideband speech is increasing in mobile communications, narrowband speech is still prevalent and therefore, a natural choice as the focus of this study. The proposed post-filter implementation fulfills also the other requirements set by the mobile communications framework, such as low delay and low computational complexity. The proposed method was compared to another post-filtering technique in terms of intelligibility in subjective tests with several different noise conditions.

## 2. Methods

A new post-filtering technique, the frequency-adaptive (FA) post-filter, was developed, and it was compared to a formant equalizing (FE) post-filtering method proposed by Hall et al. [13]. Both methods are based on the concept of transferring energy from low to high frequencies. To achieve this, FE utilizes a fixed high-pass type filter whereas FA selects an optimal filter from a collection of predefined band-stop filters. Both of the methods have low computational complexity and processing delay which makes their practical implementation in a mobile phone receiver feasible.

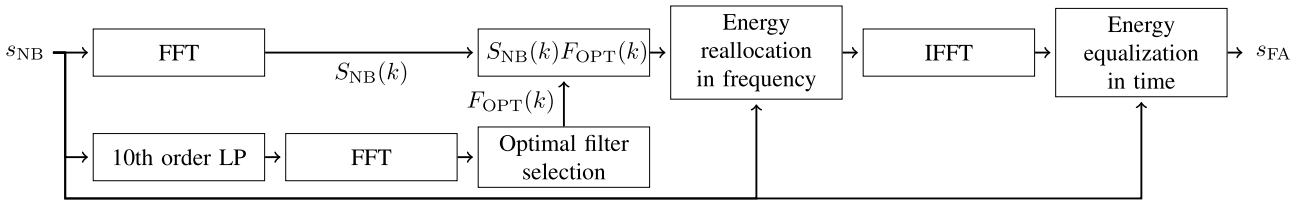


Figure 1: *Flowchart of the FA post-filtering algorithm for voiced frames. The incoming narrowband speech frame is denoted by $s_{\mathrm{NB}}$ and the processed speech frame by $s_{\mathrm{FA}}$. The fast-Fourier transform (FFT) of the speech frame and the optimal filter are denoted by $S_{\mathrm{NB}}(k)$ and $F_{\mathrm{OPT}}(k)$, respectively.*

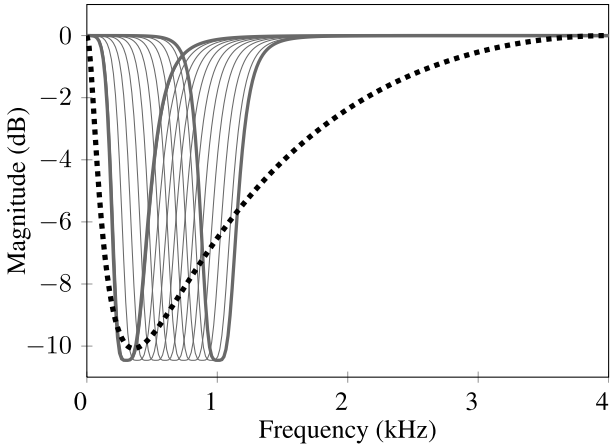


Figure 2: *The magnitude responses of the FE and FA post-filters. FE is depicted using a dotted, black line whereas a part of the filters utilized by the FA method, $F_i(k)$, are plotted using grey, solid lines.*

### 2.1. FA post-filter

The flowchart of the FA post-filtering algorithm is shown in Fig. 1. The algorithm utilizes a collection of precomputed filters, $F_i(k)$, whose stop-bands cover the approximate frequency range of the first formant (F1), as shown in Fig. 2. The filters are infinite impulse response (IIR) filters designed with the procedure given in [16]. Using the spectral envelope of the incoming speech frame, the optimal band-stop filter is selected and the filtering is done in the frequency domain. The energy removed by the filtering is then reallocated so that the energy of the processed frame is equalized to the level of the original speech frame.

The incoming speech signal is processed with a 8-kHz sampling frequency in 20-ms frames which are first windowed with $w_n = \sin(\pi/(2N) \cdot (n + 0.5))$ [17]. To reduce audible artefacts at frame borders, the same window is also applied after the processing and additionally a 50 % overlap between consecutive frames is used. The energy and the gradient-index [18] are computed from the incoming speech frame, and are used to classify the frame either as silence, unvoiced speech or voiced speech. Frames classified as silence are not processed, and for unvoiced frames, the optimal filter for the previous voiced frame followed by adaptive gain control (AGC) [19] in the time domain are used.

For voiced frames, the optimal filter is selected utilizing the spectral envelope of the frame as well as an estimate of the fundamental frequency (F0) obtained from the adaptive multi-rate (AMR) speech codec [20]. Male speakers with low F0 usually have lower F1 than female speakers [21], and therefore, the F0 estimate is used to restrict the frequency range of the filters. First, the spectral envelope is computed as the fast-Fourier transform (FFT) of the linear prediction (LP) coefficients. To determine the best band-stop filter from the predefined set of filters, the magnitude of each filter is first multiplied with the magnitude of the speech spectral envelope. Then, the outputs of the filtering operations are compared and the filter corresponding to the residual with the smallest energy is selected. The location, i.e., the center frequency of the stop-band, of the filter is compared to the location of the filter used in the previous frame. Because large changes in the filter location between consecutive voiced frames might cause audible artefacts, the final optimal filter, $F_{\mathrm{OPT}}(k)$, is a compromise between the filter removing the most energy from the current frame and the filter utilized in the previous frame if the difference between their locations in frequency is large ($\Delta f_{\mathrm{MAX}} = 50$ Hz). The distribution of the frequency locations of the optimal filter for the speech material used in the listening tests is shown in Fig. 3.

After the optimal filter has been selected, the filtering is done in the frequency domain as a multiplication of the speech spectrum, $S_{\mathrm{NB}}(k)$, qith the spectrum of the optimal filter, $F_{\mathrm{OPT}}(k)$. The energy equalization of the filtered frame to the level of the original frame is accomplished in two parts, first by reallocating a part of the energy in the frequency domain, and then by equalizing in the time domain. The purpose of this two-part equalization is to obtain a non-uniform reallocation of energy, mostly to frequency region above 1 kHz, without producing audible artefacts caused by large energy fluctuations in the 1 kHz to 4 kHz region.

The amount of energy to be reallocated in the frequency domain is computed by taking into account the energy removed by the filtering as well as the amount of energy reallocated in the previous frame ($\Delta E_{\mathrm{MAX}} = 0.25$ dB). The energy is distributed to the frequency region above 1 kHz with half a Gaussian window utilized for smoothing. Finally, the time-domain signal is obtained via inverse FFT (IFFT) and the final energy equalization to the level of the incoming speech frame is done in the time domain. A frequency-domain example of post-filtering speech with FA is shown in Fig. 4.

### 2.2. FE post-filter

The FE method introduced by Hall et al. [13] was chosen for comparison due to the similarity of the approach with the proposed post-filtering method. The algorithm utilizes a fixed high-pass filter, shown in Fig. 2, which was derived by inverting the average amplitudes of the first two formants measured from adult male speakers. The filter was originally intended for wideband speech with a 22.05-kHz sampling frequency but it was

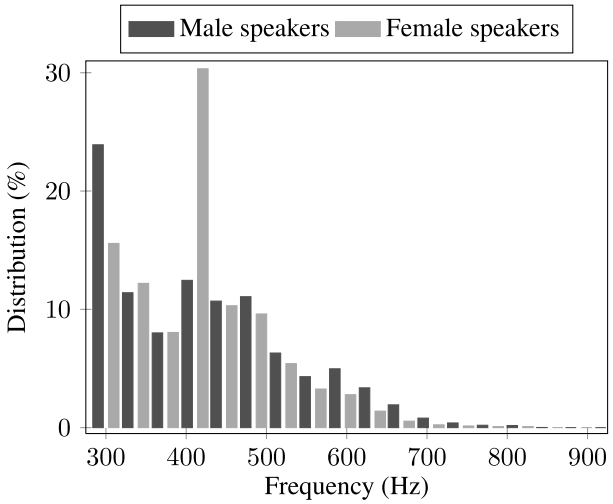


Figure 3: *The distribution of the frequency locations of the selected optimal filter, $F_{\mathrm{OPT}}(k)$, in the FA method for male and female speakers. The frequency location denotes the center frequency of the stop-band of the filter. The distributions are computed using the voiced frames in the Finnish speech material utilized in the listening tests.*

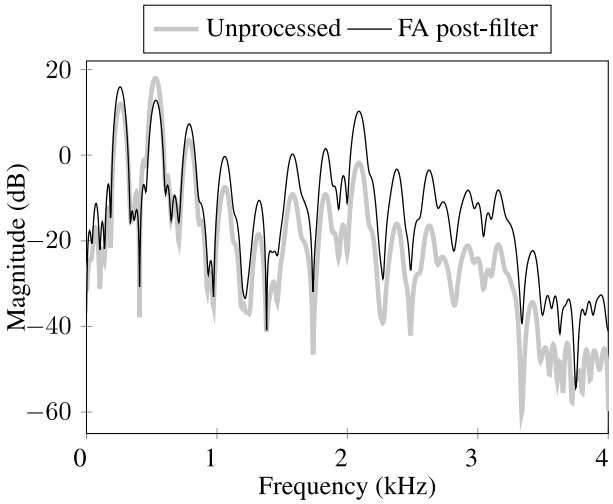


Figure 4: *A frequency domain example of post-filtering by FA (vowel /e/ produced by a female speaker)*

modified for narrowband speech using the *z* transform given in the original paper [13]. All incoming frames are processed with the post-filter, after which, the energy level of the processed frames are equalized to the level of the unprocessed frames using the AGC.

## 3. Subjective evaluation

A word-error rate (WER) test was conducted to evaluate the performance of the two post-filtering algorithms (FA and FE) in comparison to unprocessed speech (UN). In the test, clean speech was corrupted with two types of additive noise (stationary car noise and unstationary factory noise [22]). The signal-to-noise ratio (SNR) levels were selected based on informal listening. For both noise types, three SNR levels were generated to create noise conditions characterized as almost clean, moderate, and severe. For car noise, this procedure resulted in the following SNR levels: 20 dB, $-5$ dB and $-10$ dB. For factory noise, the corresponding three SNR levels were 20 dB, 0 dB, and $-5$ dB. The tests took place in a sound-proofed listening booth with Sennheiser HD 650 headphones.

The speech material consisted of phonetically balanced sentence material from four speakers (2 males) containing sentences in both Finnish and English [23]. However, only the Finnish material was used in the present test. The material has also been calibrated in terms of intelligibility in a previous study [23].

For the purposes of this test, all speech samples were preprocessed to resemble narrowband telephone speech. They were first downsampled to 16 kHz, filtered with the MSIN filter [24] to simulate mobile station input characteristics, downsampled to 8 kHz, and AMR encoded and decoded [20]. After this, the samples were equalized to $-26$ dBov with SV56 [24, 25] and post-filtered (conditions FA and FE). Finally, car or factory noise was added according to the noise condition and the noisy samples were equalized to $-26$ dBov with SV56. The samples were presented to the listeners using a 8-kHz sampling frequency.

Fifteen listeners took part in the listening test. They all evaluated their hearing to be normal and were native speakers of Finnish (except for one listener who was not a native speaker but spoke fluent Finnish). Before the actual test, a short practice session was arranged during which the listeners were able to adjust the volume to a comfortable listening level and thereafter, the volume setting was kept constant throughout the rest of the test. In the test, the subjects were allowed to play each sample only once after which they typed the sentence on the computer. The test session consisted of two parts with a short break in between and lasted approximately for one hour.

### 3.1. Results

All obvious spelling errors in the answers given by the test subjects were corrected before computing the percentage of correct words by scoring the stems and suffices of inflected words separately. The results were analyzed with a five-way analysis of variance (ANOVA) with the noise type (car noise, factory noise), the SNR level (almost clean, moderate, severe), the method (FA, FE, UN), and speaker gender (male, female) modeled as fixed variables, and the test subject as a random variable. Such a linear-mixed effects model was selected for the analysis due to the found robustness of such models against violations of the assumptions about the normality of the residuals and the normality of the random effects [26]. The ANOVA indicated that the noise type [$F(1,14) = 6.63$, $p < 0.05$], the method [$F(2,28) = 90.57$, $p \ll 0.001$], the SNR level [$F(2,28) = 545.50$, $p \ll 0.001$] as well as the interactions between the noise type and the method [$F(2,28) = 13.89$, $p < 0.001$], between the SNR level and the method [$F(4,56) = 5.96$, $p < 0.001$], and between the noise type, the SNR level and the method [$F(4,56) = 3.32$, $p < 0.05$] had a significant effect on the WER scores.

The marginal means and the 95 % confidence intervals were computed, and Dunnett's T3 post-hoc test [27] was applied to gain more insight into the nature of the effects. The values depicted in Figs. 5 and 6 illustrate that no significant differences exist between the methods at the highest SNR level, and that FA and FE achieve a significantly lower WER than UN when the SNR level is decreased. Moreover, the difference is pronounced

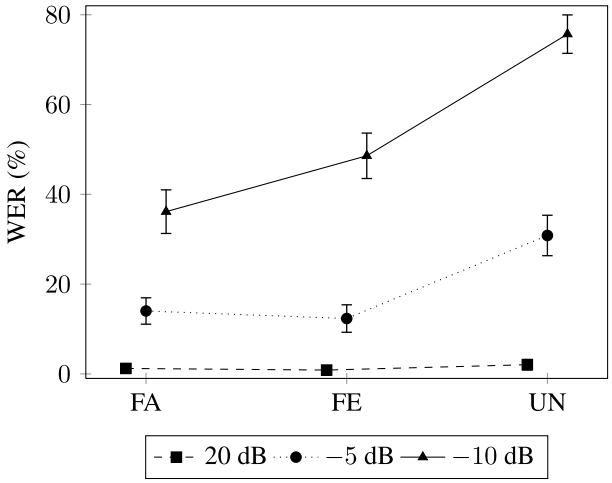


Figure 5: *Mean word-error rates (WERs) and 95% confidence intervals for car noise with three SNR values (20 dB, −5 dB, −10 dB) aggregated across male and female speakers. The three methods under comparison are the frequency-adaptive (FA) and the formant equalizing (FE) post-filters, and unprocessed speech (UN).*

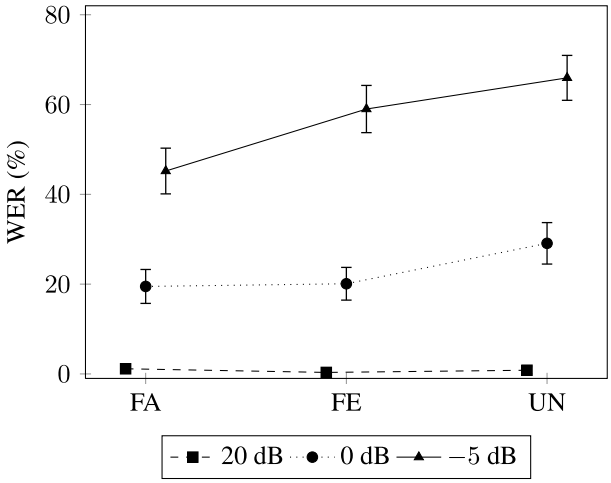


Figure 6: *Mean word-error rates (WERs) and 95% confidence intervals for factory noise with three SNR values (20 dB, 0 dB, −5 dB) aggregated across male and female speakers. The three methods under comparison are the frequency-adaptive (FA) and the formant equalizing (FE) post-filters, and unprocessed speech (UN).*

at the lowest SNR levels with the two noise types and FA yields a significantly lower WER than FE with the factory noise at −5 dB SNR level. The relationship between the WER scores obtained with FA and FE is similar also with the car noise at −10 dB SNR, although the difference is not significant in that condition. Overall, the FA method results in the lowest WER, and also the FE method yields in lower WER than the UN, as illustrated by the values depicted in Fig. 7 which contains the values computed by averaging across all of the different test scenarios. The use of the post-hoc test verified the statistical significance of the aforementioned observations.

## 4. Conclusions

A frequency-adaptive post-filtering method (FA) was introduced and compared to a similar post-filtering approach (FE) and to unprocessed speech in a subjective intelligibility test in six background noise conditions. The results indicate that both FA and FE were able to provide intelligibility improvement over unprocessed speech in adverse noise conditions. In addition, the proposed method received lower error rates on average and yielded significantly smaller error rates than FE in factory noise condition at SNR level −5 dB. With car noise, there was no significant difference in the −10 dB condition but the relationship between the post-filtering algorithms remains similar as in the case of factory noise. In accordance with the test results, informal listening indicated that the algorithm produces speech that stands out clear from the noise without becoming unnatural. However, it was also observed that some minor distortions are audible in low noise conditions (e.g. SNR $>$ 20 dB). These can be most likely avoided by restricting the fluctuations between consecutive frames.

In addition to the intelligibility improvement provided by the proposed algorithm, the method can also adapt to speaker-dependent voice characteristics in a robust manner. The adaptation does not require any training of the algorithm offline and works, therefore, equally effectively for unknown speakers as well. Furthermore, the algorithm has been designed to work in

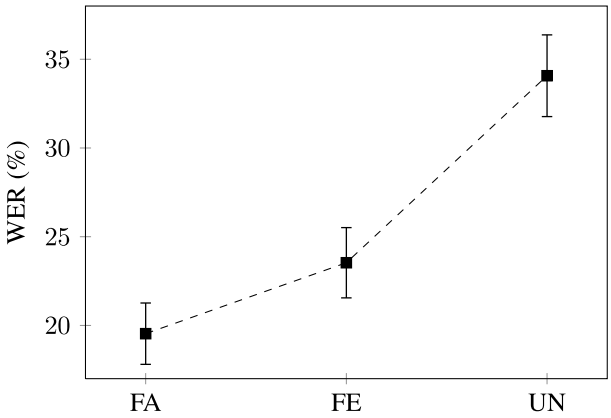


Figure 7: *Mean word-error rates (WERs) and 95% confidence intervals aggregated across all test conditions as well as male and female speakers. The three methods under comparison are the frequency-adaptive (FA) and the formant equalizing (FE) post-filters, and unprocessed speech (UN).*

mobile devices which means that it operates in real-time with minimal memory requirements.

The proposed method could be improved, for instance, by refining the pattern which is used to reallocate energy in the frequency domain. It could be adjusted to better match speaker characteristics or to emulate modifications naturally produced by speakers in adverse noise conditions. Additionally, adaptation to the level or the characteristics of the background noise could be included.

## 5. Acknowledgements

The work was supported by the GETA graduate school, Nokia, the Academy of Finland (research programmes 135003 and 13251770), and the Mide/Ui-art project of Aalto University. The authors would like to thank the listeners for help.

# 6. References


[1] J.-H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech, Audio Process.*, vol. 3, pp. 59–71, 1995.

[2] V. Grancharov, J. Plasberg, J. Samuelsson, and W. Kleijn, "Generalized postfilter for speech quality enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 57–64, 2008.

[3] H. Kim and H.-G. Kang, "An adaptive short-term postfilter based on pseudo-cepstral representation of line spectral frequencies," *Speech Commun.*, vol. 37, pp. 335–348, 2002.

[4] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. ICASSP*, 2006, pp. 493–496.

[5] ——, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, 2010.

[6] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proc. Interspeech*, 2011, pp. 345–348.

[7] ——, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech*, 2012.

[8] P. N. Petkov, W. B. Kleijn, and G. E. Henter, "Enhancing subjective speech intelligibility using a statistical model of speech," in *Proc. Interspeech*, 2012.

[9] M. Skowronski and J. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Commun.*, vol. 48, pp. 549–558, 2006.

[10] 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.

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

[12] R. Niederjohn and J. Groteluesche, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 277–282, 1976.

[13] J. Hall and J. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *J. Acoust. Soc. Amer.*, vol. 127, pp. 280–285, 2010.

[14] E. Jokinen, S. Yrttiaho, H. Pulakka, M. Vainio, and P. Alku, "Signal-to-noise ratio adaptive post-filtering method for intelligibility enhancement of telephone speech," *J. Acoust. Soc. Amer.*, vol. 132, no. 6, pp. 3990–4001, 2012.

[15] V. Hazan and A. Simpson, "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," *Speech Commun.*, vol. 24, pp. 211–226, 1998.

[16] M. Holters and U. Zölzer, "Parametric higher-order shelving filters," in *Proc. EUSIPCO*, 2006.

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

[18] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Process.*, vol. 86, pp. 1296–1306, 2006.

[19] 3rd Generation Partnership Project (3GPP), "Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions, 3GPP TS 26.090," 2008, version 8.0.0.

[20] 3rd Generation Partnership Project (3GPP), "ANSI-C code for the floating-point adaptive multi-rate (AMR) speech codec, 3GPP TS 26.104," 2009, version 9.0.0.

[21] D. G. Childers and K. Wu, "Gender recognition from speech. Part II: Fine analysis," *J. Acoust. Soc. Amer.*, vol. 90, no. 4, pp. 1841–1856, 1991.

[22] A. Varga and H. J. 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 Commun.*, vol. 12, pp. 247–251, 1993.

[23] M. Vainio, A. Suni, H. Järveläinen, J. Järvikivi, and V.-V. Mattila, "Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish," *J. Acoust. Soc. Amer.*, vol. 118, pp. 1742–1750, 2005.

[24] ITU-T, "Recommendation G.191 : Software tools for speech and audio coding standardization," 2005.

[25] ——, "Recommendation P.56 : Objective measurement of active speech level," 1993.

[26] H. Jacqmin-Gadda, S. Sibillot, C. Proust, J.-M. Molina, and R. Thiébaut, "Robustness of the linear mixed model to misspecified error distribution," *Comput. Stat. Data Anal.*, vol. 51, no. 10, pp. 5142–5154, 2007.

[27] C. W. Dunnett, "Pairwise multiple comparisons in the unequal variance case," *J. Amer. Statist. Assoc.*, vol. 75, no. 372, pp. 796–800, 1980.