COMPARISON OF POST-FILTERING METHODS FOR INTELLIGIBILITY ENHANCEMENT OF TELEPHONE SPEECH

Emma Jokinen[1], Paavo Alku[1], Martti Vainio[2]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland
[2]Institute of Behavioural Sciences, University of Helsinki, Finland
emma.jokinen@aalto.fi

ABSTRACT

Post-filtering can be used to enhance the quality and intelligibility of speech in mobile phones. This paper introduces two straightforward post-filtering methods for near-end speech enhancement in difficult noise conditions. Both of the algorithms use a perceptually motivated high-pass filter to transfer energy from the first formant to higher frequencies. A Speech Reception Threshold (SRT) test was conducted to determine the performance of the proposed methods in comparison to a similar post-filtering approach and unprocessed speech. The results of the listening tests indicate that all of the post-filtering methods provide intelligibility enhancement compared to unprocessed speech, but there were no significant differences between the methods themselves.

***Index Terms*—** Speech processing, speech enhancement, post-filtering, intelligibility, SRT

1. INTRODUCTION

In speech communication technology, post-processing refers to methods to enhance the quality and intelligibility of speech that has been degraded by acoustical background noise and by quantization noise generated by low bit-rate speech coders. In mobile phones, post-processing is typically implemented in the form of *post-filtering*. This refers to filtering the decoded speech signal at the mobile phone receiver with an adaptive filter in order to reduce the effects of quantization noise thus enhancing the perceptual quality of speech. The enhancement is typically achieved by emphasizing the spectral peaks and attenuating the noise components at spectral valleys, a procedure that is taken advantage of, for example, by the widely used post-filtering method introduced in [1]. The technique proposed in [1] has since been modified in several studies by utilizing, for example, a modified Yule-Walker filter [2] and psychoacoustic models [3]. In addition, a generalized version of the post-filter was introduced in [4] in order to better adapt to a large variety of noise conditions. However, most of the known post-filtering methods have been designed for and tested in high to moderate signal-to-noise ratios (SNRs) where the quality of speech is the main concern.

In mobile communications, phone users are often in situations where improving the speech quality alone is not sufficient. In adverse noise conditions, the intelligibility of speech is severely compromised and becomes the most important factor to be preserved in order to enable conversation between phone users. Several post-processing methods have therefore been proposed to address the improvement of speech intelligibility in noise. In [5], for example, selective boosting is used to reallocate speech energy to frequency regions with low SNRs. In [6], a similar approach is utilized by transferring energy from voiced sounds to unvoiced sounds. The idea of energy reallocation has also been used in post-filtering algorithms ([7, 8]), where high-pass type post-filters are used to attenuate low frequency regions and to enhance higher frequencies, effectively transferring more energy to upper frequencies. Results achieved in [6]-[8] indicate that the reallocation of speech energy with high-pass filtering can be used to produce more intelligible speech. It is, though, worth noting that the results of [6]-[8] have been obtained with test material that consists of individual words. Even though these previous studies have undoubtedly provided valuable information about speech intelligibility in noise, the experiments conducted have been unable to simulate realistic speech communication situations encountered by mobile phone users.

It is highly challenging to evaluate speech intelligibility in realistic conditions that correspond well to the every-day use of mobile phones. In real situations, the speech signal is continuous and therefore using short, individual samples does not provide an accurate estimate of intelligibility [9]. The SRT test aims to solve these problems by using an adaptive procedure on a list of samples to determine the intelligibility score. In the SRT test, the first sample of a list is played on a predetermined SNR level and the listener is asked to type in what he or she heard. The answer is compared to the correct one and the presentation level of the following sample is adjusted based on how much the listener was able to understand. This procedure is repeated throughout the list of samples to attain the SNR level where a certain percentage of speech is understood. Thus, the SRT test provides an accurate means of estimating speech intelligibility and avoids floor and ceiling effects [10].

This study addresses the utilization of post-filtering in improving the intelligibility of narrowband telephone speech which has been encoded according to speech coders used in mobile phones. Two straightforward post-filtering algorithms are proposed and their behavior is compared with a reference technique in an SRT test in two realistic noise conditions. To the knowledge of the authors, post-filtering methods have not previously been tested with the SRT test. The results of the SRT evaluation indicate that post-filtering improves the intelligibility compared to unprocessed speech and that the performance is similar between the algorithms tested.

2. METHODS

Three post-filtering methods were implemented with MATLAB to be evaluated in the SRT test. Two of these algorithms, adaptive (AD) post-filter and fixed (FI) post-filter, are new and they were specifically designed for the SRT experiments of this work. The third one (FE) is a reference method recently proposed by Hall *et al.* [8]. All of these three post-filtering algorithms are based on the general idea of reallocating energy from low frequency regions to higher frequencies. In addition, they all share such general properties (*e.g.*, computational complexity and delay) that make, in principle, the algorithms implementable in mobile phone receivers.

2.1. AD post-filter

The flowchart of the AD post-filtering algorithm is depicted in Fig. 1. The incoming speech signal, s_{NB}, is processed in 20-ms frames with 8-kHz sampling frequency. The frames are classified as voiced, unvoiced or silence using the gradient-index (GI) [11] and the energy of the pre-emphasized frame. The pre-emphasis is done with $H(z) = 1 + \mu z^{-1}$ where μ is the first linear prediction (LP) coefficient of the speech frame. Silent frames are not processed and unvoiced frames are processed only if the next frame is voiced. Thus, some information on the next frame (*i.e.*, gradient-index and frame energy) is also required. If the next frame is voiced, the transition between the frames is smoothed. For voiced frames, the formant frequencies are estimated by combining conventional post-filtering [1] with spectral peak-picking. The first ten LP coefficients of the frame are used to form a short-term post-filter structure which has more dominant formant peaks. The first three peaks are located from the spectrum and the two peaks which are closest to the formants of the previous frame are chosen as the first two formants.

The post-filter consists of three filters in cascade. The transfer functions of the first two filters are given as

$$H_i(z) = \frac{1 - 2 \cdot 0.9 \cdot \cos(\theta_i) \cdot z^{-1} + 0.9^2 \cdot z^{-2}}{1 - 2 \cdot r_i \cdot \cos(\theta_i) \cdot z^{-1} + r_i^2 \cdot z^{-2}}, i = 1, 2 \tag{1}$$

where θ_i's denote the locations of the first and second formant and the parameters r_i are chosen so that the first formant is attenuated, $0 < r_1 \leq 0.9$, and the second formant is enhanced, $0.9 \leq r_2 < 1$. The value 0.9 was fixed through informal listening to constrain the number of free parameters. The final component of the post-filter structure is a first-order low-pass filter which is used to compensate for the possible tilt resulting from the cascade of the two formant filters. The coefficient of this tilt-filter is determined from a first-order LP analysis on the cascade of the formant filters.

In order to avoid sudden transitions, the coefficients of the post-filter are linearly interpolated in the line spectral frequency domain between two consecutive voiced frames. After filtering, the overall energy of the processed frame is set to the level that is equal to that of the original frame using the adaptive gain control (AGC) in the AMR-NB standard [12].

The parameters of the post-filter, r_1 and r_2, were determined by conducting informal listening tests. In these tests, 17 listeners, all native speakers of Finnish, were asked to choose their preferred parameter values for six Finnish speech samples corrupted with car noise (SNR -5 dB). The purpose of the test was to obtain parameter values yielding speech which is intelligible and clear but pleasant to listen to. The listeners commented that the samples became more intelligible and clear when the first formant was strongly attenuated ($r_1 < 0.5$). The sharpening of the second formant peak was often found to sound annoying and to result in a strong whistling effect ($r_2 < 0.96$). The average values were chosen for the post-filter.

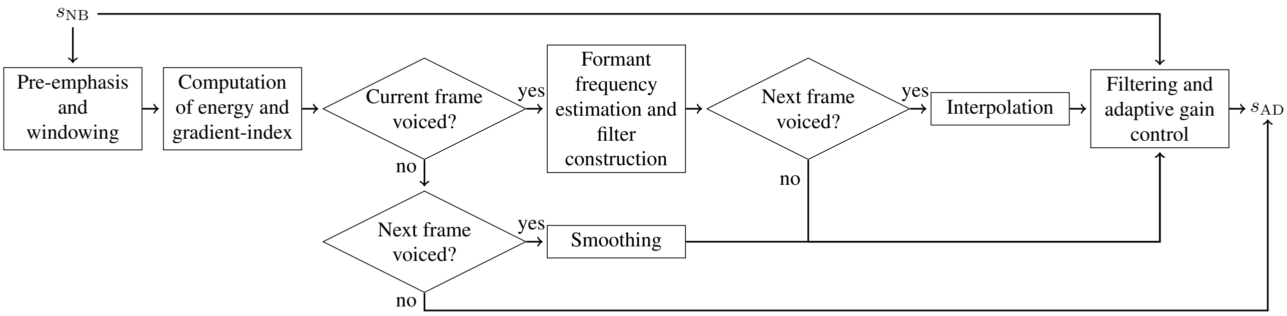
2.2. FI post-filter

The FI post-filter uses the same filter structure as the AD method. The only difference is that the post-filter does not adapt its coefficients based on formant locations but uses constant values of θ_i and consequently, the coefficient of the tilt-filter also remains unchanged. Suitable values of θ_i for describing the approximate average formant locations were determined by processing 400 Finnish sentences [9] with the AD algorithm and using the averages of the estimated formant frequencies, $F_1 = 500$ Hz and $F_2 = 1600$ Hz. The required tilt correction was computed in the same way as for the AD post-filter.

All speech frames are processed with the FI post-filter. After filtering, AGC is used to equalize the energies of the processed and original speech frames.

2.3. FE post-filter

The FE post-filter introduced by Hall *et al.* [8] was chosen for comparison due to its similarity with the methods proposed in this paper. The FE method utilizes a fixed post-filter, which is designed by inverting the amplitudes of the first two formants of adult male speakers resulting in a high-pass type filter. The



s_{NB}
Pre-emphasis and windowing
Computation of energy and gradient-index
Current frame voiced?
yes
no
Formant frequency estimation and filter construction
Next frame voiced?
yes
no
Interpolation
Filtering and adaptive gain control
s_{AD}
Next frame voiced?
yes
Smoothing
no

Fig. 1. Flowchart of the AD post-filtering algorithm. The input speech signal is denoted by s_{NB} and the processed signal is s_{AD}.

post-filter was originally intended for wideband speech with 22.05 kHz sampling rate but it was adapted to narrowband by using the z-transform given in the appendix of [8]. The post-filter is used for all speech frames and after the filtering, the energy levels of the processed frames are equalized with the original ones using AGC.

3. SUBJECTIVE EVALUATION

An SRT test was conducted to compare the intelligibility improvement given by the developed post-filtering methods and the reference method to that of unprocessed speech. In the SRT test, the presentation level (in terms of SNR) of the next sample is adjusted according to the listener's typed response to the previous sample. The purpose of the adaptive procedure is to determine the SNR level where the listener understands 50 % of the speech samples.

3.1. Speech material

The sentence material used in the SRT test was developed by Vainio *et al.* [9]. The Finnish material consists of 25 lists of 16 phonetically balanced sentences spoken by two female and two male speakers. The first 24 lists were selected for the actual test and list 25 was reserved for demonstration purposes.

The test samples were first downsampled to 16 kHz and filtered with the MSIN filter [13] to simulate mobile phone input characteristics. Next the speech signals were downsampled to 8 kHz and AMR coded and decoded twice using the C-implementation provided by 3GPP [14]. The encoding and decoding was done twice to take into account the possibility of multiple encodings/decodings in a real system. Then the samples were equalized to -26 dBov with SV56 [13, 15] at 16 kHz, downsampled back to 8 kHz and processed with one of the post-filters, AD, FI or FE. In case of the unprocessed reference condition (UN), no post-filtering was used. After this, the samples were once again upsampled to 16 kHz and equalized with SV56. Finally, speech-shaped noise or car noise was added depending on the test condition. The speech-shaped noise was generated based on the long term average spectrum of the Finnish SRT speech material [9] and was chosen because it was specifically designed for the SRT test. Car noise, which was stationary, low-pass type noise, was included to obtain a more realistic test condition. The noisy samples were equalized to the same level with SV56 before being played to the listeners.

3.2. Test procedure

There were 17 normal-hearing listeners between ages 22 and 41 participating in the SRT test. For the speech-shaped noise condition, 12 listeners were used and the remaining 5 listeners took part in the car noise condition. All of the participants were native speakers of Finnish and either students or staff at the Aalto University. The listening tests were conducted in a quiet office space with Sennheiser HDA 200 headphones and a laptop. In the beginning of the test session, the listener was given written instructions which for instance advised to pay special attention to spelling in their responses. Before the actual test, there was a short practice session during which the listener was allowed to ask questions and to adjust the volume setting to a comfortable listening level. After this, the volume setting was kept constant. The actual test was split into three parts separated by short breaks. One test session took approximately one hour to complete.

One test consisted of 16 lists with 10 samples each. All of the samples in a single list were spoken by the same person and were processed with the same post-filtering method. Each method under comparison was paired once with each speaker. The selection of the lists and the samples was random as was the presentation order of the lists and the processing methods. The only restriction was that all of the lists and samples of one listener were different thus avoiding possible learning effects.

The first sample of each list was at first played with SNR level -10 dB. The percentage of correct words in the typed answer was checked using the guidelines in [9]. For instance, both the present and imperfect tense of a verb were accepted since they are very similar in Finnish. In addition, if the ending of an inflected word was incorrect, half a word point was given for the correct word stem. The presence of extra words did not matter but the correct words had to be in the right order. The SNR level of the next sample on the list was adjusted based on the listener's answer according to the approach suggested in [16] while the level of the background noise was kept constant. The first sample of a list was played with an increasing SNR level until the subject understood over 60 %. All of the other samples in the list were played only once. The SRT value for a list was determined by calculating the average presentation level of samples 5–11 on the list. Even though there are only 10 samples, the presentation level of the 11th sample can be determined from the response given to the last sample.

3.3. Results

The SRT test is sensitive to spelling errors because they misdirect the adaptive adjustment of the SNR. Even though the listeners were instructed to pay special attention to correct spelling, some errors were still made. Therefore, lists that contained clear spelling errors were discarded from the results. In addition, one list where the SNR after the first sentence was 33 dB was eliminated as an outlier. For each listener, the two lists with the same processing and speaker gender were averaged together. If for some condition one of the lists had been discarded, the average was replaced with the score of the remaining list. The listeners who had errors in both lists for one condition were discarded altogether. As a result, two listeners in the speech-shaped noise condition and one listener in the car noise condition were discarded from the final results. The average SRT scores computed from the remaining listeners for both noise types are depicted in Fig. 2.

A repeated measures ANOVA with two factors, method (UN, AD, FI, FE) and speaker gender (male, female), and one between-subjects factor, noise type (speech-shaped, car) was conducted on the mean values of the SRT scores. The analysis showed that the effects of method $[F(3, 36) = 44.54, p < 0.001]$, gender $[F(1, 12) = 74.27, p < 0.001]$ and noise type $[F(1, 12) = 72.61, p < 0.001]$ were statistically significant. Also, the interaction term method×gender, $[F(3, 36) = 2.65, p < 0.05]$, was significant. Post hoc tests with the Bonferroni adjustments indicated that all of the post-filtering methods performed better than the unprocessed reference condition (p values < 0.001). However, there were no statistically significant differences between AD, FI and FE. The analysis also showed that SRT scores for male speakers (SNR -6.64 dB) were better than for female speakers (SNR $-$ 4.65 dB) ($p < 0.001$) and that speech-shaped noise (SNR $-$ 3.21 dB) was more difficult than car noise (SNR $-$ 8.08 dB) ($p < 0.001$).

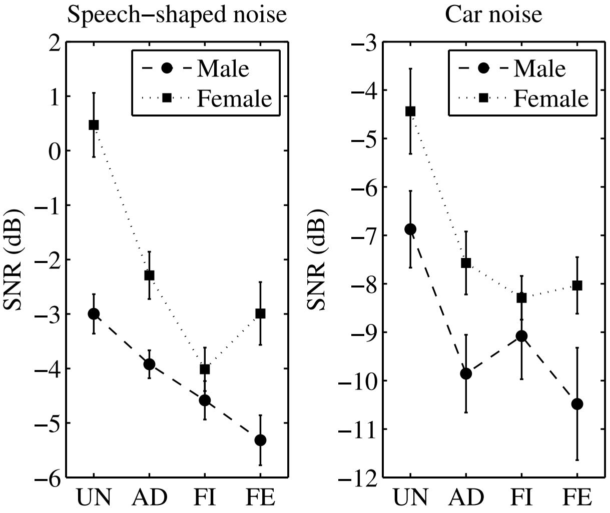


Fig. 2. Results of the SRT tests for both noise types and for male and female speakers. The four methods under comparison are on the x-axis and they correspond to unprocessed speech (UN), the adaptive (AD) post-filter, the fixed (FI) post-filter, and the formant equalizing (FE) post-filter. The markers represent the average SRT scores (SNR levels required for 50 % intelligibility) of the methods for different conditions. Lower scores indicate better enhancement performance. The error bars denote the standard errors of the mean.

4. DISCUSSION

Two post-filtering methods (AD and FI) were introduced and compared to a similar post-filtering approach (FE) and to unprocessed speech in an SRT test in two background noise conditions. The results indicate that all post-filtering methods evaluated were able to provide intelligibility improvement over unprocessed speech in adverse noise conditions. However, there were no statistically significant differences between the performance of the three algorithms and it can, thus, be concluded that the adaptivity did not provide any additional benefit in terms of intelligibility. Both of the non-adaptive methods, FI and FE, have a small computational load, whereas the AD post-filter requires more computation in searching the formants, adjusting the filter accordingly and interpolating between consecutive frames. This could be justified with possible improvements in quality, but in informal listening the differences between the AD and FI post-filters were almost inaudible. One reason for this is the simplicity of the formant frequency estimation in the AD algorithm which results in errors in the locations of the formants. Therefore, the AD post-filter does not necessarily suppress the first formant or amplify the second formant of each frame exactly, but some frequency regions near them. As a result, the AD and FI methods are very similar, especially since the structure of the post-filter is simple and robust.

The statistical analysis also suggests that there are significant differences between the intelligibility of processed speech between male and female speakers. The post-filtering methods used here affect male voices more probably because male speakers tend to have lower fundamental frequencies and thus more energy in the low frequency region. The more prominent effect with male speakers was also noticed in the informal subjective tests that were used in optimizing the parameters of the formant filters. A few listeners commented that the processing changed the tone of the speakers and they found this slightly disturbing especially with the male voices. For female speakers, the change in tone was not as noticeable.

The SRT test had some issues which might have affected the statistical power of the test. First, the sensitivity of the adaptive procedure to spelling errors resulted in the elimination of three listeners from the final results. Clearly, some elementary logic should be implemented to the automated checking procedure to prevent this. Second, there were large deviations in the average scores obtained from the test. Differences in individual listeners' hearing cause some variation but there was also large intra individual variation present. The test results could possibly be further improved by computing the SRT score of a list from the SNR levels of only the last 4 or 5 samples. As the adaptive procedure continuously approaches the correct SNR estimate, taking the average from fewer samples would presumably provide a more accurate estimate.

5. ACKNOWLEDGEMENTS

The study was supported by the GETA graduate school and Nokia. The authors would like to thank the volunteers who participated in the listening tests.

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] A. Mustapha and S. Yeldener, "An adaptive postfiltering technique based on the modified Yule-Walker filter," in *Proc. ICASSP*, 1999, pp. 197–200.

[3] W. Chen, P. Kabal, and T.Z. Shabestary, "Perceptual postfilter estimation for low bit rate speech coders using Gaussian mixture models," in *Proc. Interspeech*, 2005, pp. 3161–3164.

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

[5] 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.

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

[7] R.J. Niederjohn and J.H. Grotelueschen, "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.

[8] 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, pp. 280–285, 2010.

[9] 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.

[10] M. Nilsson, S.D. Soli, and J.A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Amer.*, vol. 95, pp. 1085–1099, 1994.

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

[12] 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.

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

[14] 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.

[15] ITU-T, "Recommendation P.56 : Objective measurement of active speech level," 1993.

[16] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," *J. Acoust. Soc. Amer.*, vol. 111, pp. 2801–2810, 2002.