INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Intelligibility enhancement at the receiving end of the speech transmission system — effects of far-end noise reduction

Emma Jokinen and Paavo Alku

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

Abstract

Post-processing methods can be used in mobile communications to improve the intelligibility of speech in adverse near-end background noise conditions. Generally, it is assumed that the input of the post-processing contains quantization noise only, that is to say, no far-end noise is present. However, this assumption is not entirely realistic. Therefore, the effect of far-end noise with and without noise reduction on the performance of three post-processing methods is studied in this investigation. The performance evaluation is done using subjective intelligibility and quality tests in several far-end and near-end noise conditions. The results suggest that although the noise reduction generally improves performance in stationary far-end noise, the noise reduction does not improve intelligibility in unstationary far-end noise conditions but has a positive impact on perceptual quality for some of the post-processing methods.
Index Terms: post-processing, intelligibility enhancement, word-error rate, far-end noise, spectral subtraction

1. Introduction

In mobile communications, post-processing can be used to enhance the intelligibility of speech in the presence of background noise in the listener's environment. This is referred to as the near-end noise scenario and it is usually assumed that the decoded speech is distorted by quantization noise only. This means that the sending side of the telephone connection, referred to as the far-end side, is assumed to be free from background noise. In this scenario, the post-processing aims to enhance the acoustic cues in the clean speech signal to improve its intelligibility over the background noise in the near-end.

Several methods of intelligibility enhancement have been proposed for the near-end condition. They are based, for example, on optimizing objective measures, such as the speech intelligibility index (SII) [1, 2, 3] and the glimpse proportion [4], or re-allocating speech energy with simple high-pass filtering [5, 6]. It is worth noting that most of these techniques have been developed by using, in one form or another, models of human speech perception. Methods to improve intelligibility based on modeling the human speech production mechanism have also been proposed. For instance, the Lombard effect [7] has been used in previous post-processing studies [8, 9, 10], and recently, Gaussian mixture models (GMMs) were successfully used in a normal-to-Lombard mapping to improve the intelligibility of telephone speech [11, 12].

The assumption of clean far-end conditions that has been made in most of the intelligibility enhancement studies is unrealistic. If the input signal contains noise, the post-processing might modify and enhance the noise along with the signal which might result in quality and intelligibility degradation. This was observed in [13], when studying intelligibility using an objective intelligibility metric, the extended speech intelligibility index (eSII). One solution is to combine a pre-processing stage suppressing the far-end noise in the sending device with the post-processing stage in the receiving device. Although, noise reduction can increase the perceptual quality of speech in stationary noise conditions [14], it can also produce artifacts to the processed signal [15] which might further degrade the post-processed speech. The authors of [13] also combined the post-processing method with noise reduction and demonstrated increased eSII scores with far-end noise. However, only stationary speech-shaped noise was used in the far-end and the objective results were not verified using subjective tests.

This study investigates the effects of having corrupted speech inputs to post-processing methods at the receiving end of the speech transmission system. In the analyzed scenario, the corruption in the input of the post-processing algorithms results from far-end environmental noise which is either processed or not processed with a noise reduction algorithm at the transmitting end. The subjective performance evaluation consists of an intelligibility test with two levels of near-end noise as well as a quality test, both conducted with several types of far-end noise. Three different post-processing methods are included in the evaluation together with one spectral subtraction method that can be used as a pre-processing step to reduce the far-end noise. All the evaluations are conducted using narrowband telephone speech.

2. Methods

Three different types of algorithms were used for post-processing: formant enhancement with fixed high-pass filtering (FE), Lombard-tilt modeling using GMMs (LTM) and dynamic range compression (DRC). All of the algorithms were operated under the normal assumption of clean speech input, that is to say they were not adapted to the presence of the far-end noise in any way. In addition to the post-processing methods, spectral subtraction (SS) was used in the noisy far-end conditions as a pre-processing stage to reduce the noise. The different stages of signal processing and the use of the different algorithms are depicted in Fig. 1. All of the algorithms are described in more detail in the following.

2.1. Formant enhancement (FE)

The FE method was introduced by Hall and Flanagan [5]. The algorithm utilizes a fixed high-pass filter which was derived by inverting the average amplitudes of the first two formants measured from adult male speakers. The resulting filter attenuates the frequency range around the first formant with maximum attenuation near 360 Hz. The filter was originally intended

http://dx.doi.org/10.21437/Interspeech.2016-144

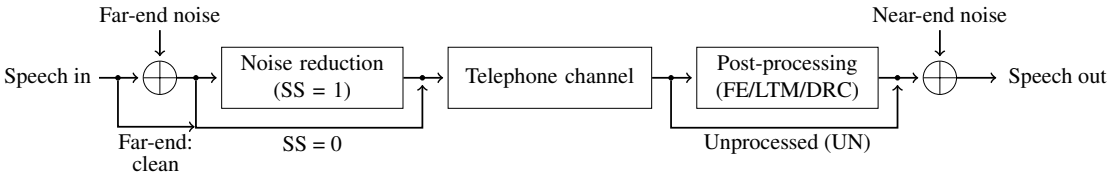

Figure 1: An illustration of how the far-end and near-end noises are added to the speech signal and where the different processing steps are done. The spectral subtraction (SS) is used before the telephone channel whereas the post-processing methods (FE, LTM and DRC) are utilized after the speech has been received from the channel and decoded. In conditions where SS = 0, the noise reduction stage is bypassed. Similarly, in the unprocessed (UN) condition, the post-processing step is omitted.

for wideband speech with a 22.05-kHz sampling frequency but it was modified for narrowband speech using the z transform given in the original paper [5].

2.2. Lombard-tilt modeling (LTM)

The LTM method, first introduced in [11], aims to map the spectral tilt from normal speech to that of Lombard speech using GMMs. The normal-to-Lombard mapping is trained using read speech data in Finnish from three male and three female speakers. The data contains parallel recordings of normal and Lombard speech, that is to say, each sentence has been produced in both conditions by each speaker. To find corresponding normal and Lombard frames, the samples are aligned using dynamic time warping (DTW) [16]. As the Lombard effect is not consistent throughout the recordings, the data selection scheme introduced in [17] was used to select training data. The trained mapping is then utilized as a part of a post-processing algorithm which has been previously used in [11].

First, the spectral tilt, parametrized as $1/A_{\mathrm{p}}(z)$, is estimated with stabilized weighted linear prediction [18], transformed to the LSF representation and mapped with the trained model. After the mapping, the stability of the output filter is checked and if necessary, the roots outside of the unit circle are replaced with their mirror-image pairs inside the unit circle. The speech frame is then filtered with $A_{\mathrm{p}}(z)/A'_{\mathrm{p}}(z)$ which removes the original spectral tilt and replaces it with the Lombard-like spectral tilt. Finally, the energy of the filtered frame is equalized to the level of the unprocessed frame with the adaptive gain control (AGC) used in the AMR codec [19].

2.3. Dynamic range compression (DRC)

The DRC method utilized is based on [8]. An slightly modified version optimized to noisy far-end conditions was proposed in [13] but this was not utilized in the current study. The compression is done in two stages: a dynamic stage and a static stage. In the dynamic stage, the estimated envelope of the signal is smoothed utilizing attack and release time constants adapted to the lower sampling rate of 8 kHz from 16 kHz using the definition given in [20]. In the static stage, a time varying gain is determined based on the decibel value of the smoothed envelope and the input-output envelope characteristic function. The 0 dB reference level needed to determine the decibel value of the envelope was set to 30% of the maximum envelope of the speech database used in the current study.

In their original study [8], the authors use a sentence level energy normalization after the compression. For the purposes of the current study, the method was adapted to real-time processing by implementing it in a frame-based form. Originally, frame-based processing was used only to compute the envelope of the original sample with different frame lengths for male and female speakers. The frame length used in this study was selected as a compromise between these and was set to 15 ms.

2.4. Spectral subtraction (SS)

The SS algorithm is based on Wiener filtering [15] and the noise power is estimated using the method introduced in [21]. The implementation was done using the Matlab toolbox Voicebox [22]. Several parameter options were evaluated informally and finally, the Wiener filter method was found to provide the best quality output speech in the conditions that were used in the study. The frame length for the noise reduction was 16 ms.

3. Subjective evaluation

The post-processing methods (FE, LTM and DRC) were evaluated in comparison to unprocessed speech (UN) in different combinations of far-end and near-end noise conditions with and without spectral subtraction using two types of subjective tests. The first test was a word-error rate (WER) test evaluating the intelligibility of the methods in the different conditions. The second test was a pair comparison test on quality.

An overall view of the different conditions used for processing can be seen in Fig. 1. Three far-end noise conditions were evaluated: clean condition as well as two noisy conditions with car noise and factory noise [23]. The levels of the far-end noise were selected outside of the normal signal-to-noise ratio (SNR) range in which they are tested [24]. This was done in order to focus on challenging conditions where the performance of the noise reduction deteriorates. Therefore, for the car noise the SNR was set to 0 dB and for the more difficult, unstationary factory noise to 5 dB. The far-end noise conditions were the same in both subjective tests.

In order to restrict the number of conditions in the evaluation, only one type of near-end noise, stationary car noise, was used. However, in the WER test, two different SNR levels, characterized perceptually as moderate and severe, were utilized. For clean far-end conditions, the SNRs for the two conditions were set to -5 dB (moderate) and -10 dB (severe), whereas for the noisy far-end conditions, perceptually similar noise levels were achieved when the SNR levels were 0 dB (moderate) and -5 dB (severe). In the pair comparison test, only one near-end noise level was used with SNR 5 dB. The near-end SNR was increased from the WER test to obtain conditions where the possible artifacts caused by the different stages of processing would not be entirely masked but the level of noise would still merit the use of intelligibility enhancement to facilitate the understanding of the samples.

The Finnish speech material used in the test consisted of phonetically balanced sentence material from two male and two female speakers [25]. The sentences used in the material are

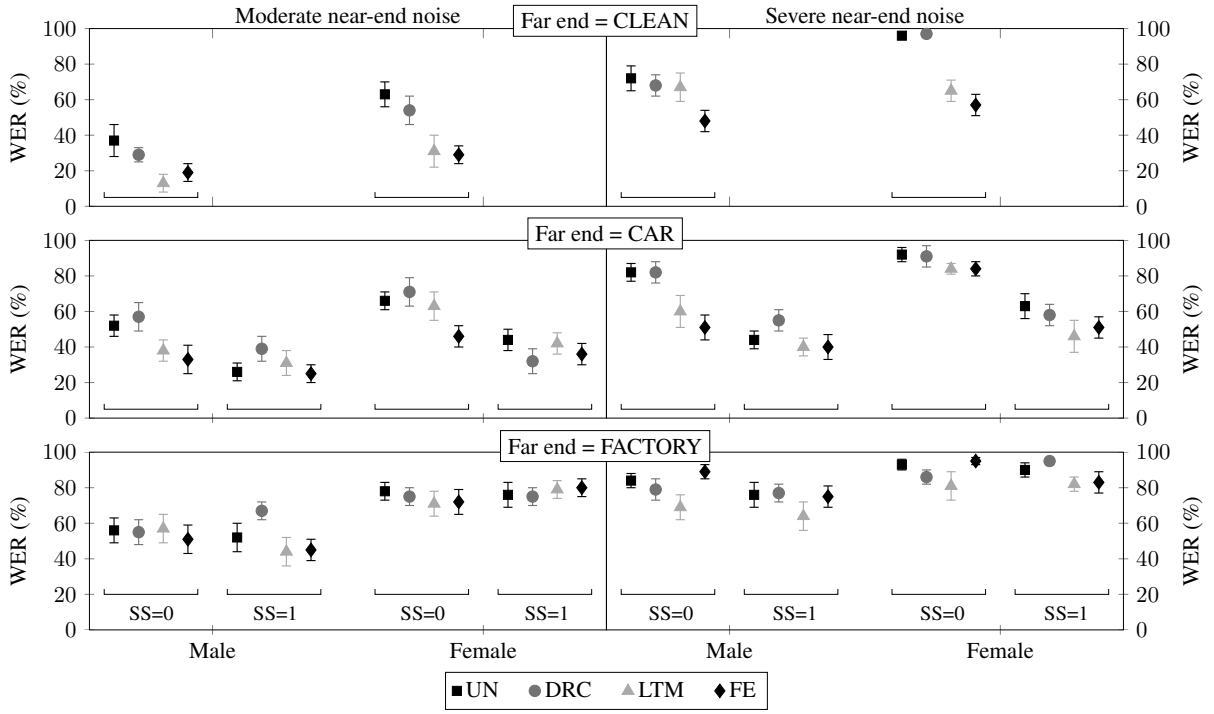

Figure 2: The mean word-error rates (WERs) and the standard errors of the mean in all test conditions for both male and female speakers. The post-processing methods under comparison were unprocessed speech (UN), dynamic range compression (DRC), Lombard-tilt modelling (LTM) and the formant equalizing postfilter (FE). Spectral subtraction was used as a pre-processing step in the cases marked with SS=1. On the top row, the clean far-end noise condition is shown. The far-end conditions with car and factory noise are depicted in the middle and bottom rows, respectively. The moderate and severe near-end noise conditions are shown on the left and right column, respectively.

semantically meaningful and have an average duration of approximately 2 seconds. All of the processing was done using the guidelines given in [26] for narrowband speech. In the conditions with far-end noise, both the speech and noise file were separately filtered with the MSIN filter [27] at 16 kHz, downsampled to 8 kHz and adjusted to the desired level using SV56 [28]. In practice, the speech level was set to -26 dBov and the noise level was corrected based on the desired SNR. After this, the far-end noise and speech were added together in the noisy far-end conditions. After this, spectral subtraction was used where appropriate. In the noisy far-end conditions, all the post-processing methods as well as UN were evaluated both with a pre-processing step consisting of noise suppression and without it. In the clean far-end condition, no noise was added to the speech. The far-end signal was then encoded and decoded using the adaptive multi-rate (AMR) narrowband codec with 12.2-kbit/s rate [29, 19]. After the decoding, the samples were processed using one of the post-processing methods under comparison except in the unprocessed condition. Based on the desired near-end SNR, the level of the near-end noise signal was adjusted and the noise signal was then added to the processed speech signal. After adding the noise, the signals were again normalized to -26 dBov using SV56 [28].

For both the WER test and the pair comparison test, eleven normal-hearing listeners took part. All subjects were native or bilingual speakers of Finnish except for one listener who was non-native but spoke fluent Finnish. In the WER test, the average age of the listeners was 24 years whereas in the pair comparison test the average was 27 years. The listeners were either university staff or students and they were paid for their participation. The tests took place in a sound-proofed listening booth using Sennheiser HD 650 headphones. Both subjective tests were divided into three parts according to the far-end noise condition with a short break in between. In the WER test, a short practice session preceded each part to acquaint the listeners to the change in the background noise. For the pair comparison, a single practice session was completed before the beginning of the test. The A-weighted sound pressure level was set to approximately 70 dB and kept constant throughout the tests.

In the WER test, each noisy sample was played only once after which the subjects typed the sentence on the computer. The percentage of correct words was computed by scoring the stems and suffices of inflected words separately after obvious spelling errors had been manually corrected. In the pair comparison test, the listeners were able to freely listen to two samples, A and B, and were asked *"Which sample is of better quality?"*. 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.

3.1. Results

For the pair comparison test, a summary score on quality was computed for each method by averaging the number of comparisons the method was preferred in each condition per listener. The results were aggregated across male and female speakers. For both the WER and the pair comparison test, the statistical analysis of the scores was divided into three separate parts according to the far-end noise condition. The WER scores were analyzed using analysis of variance (ANOVA) with the method

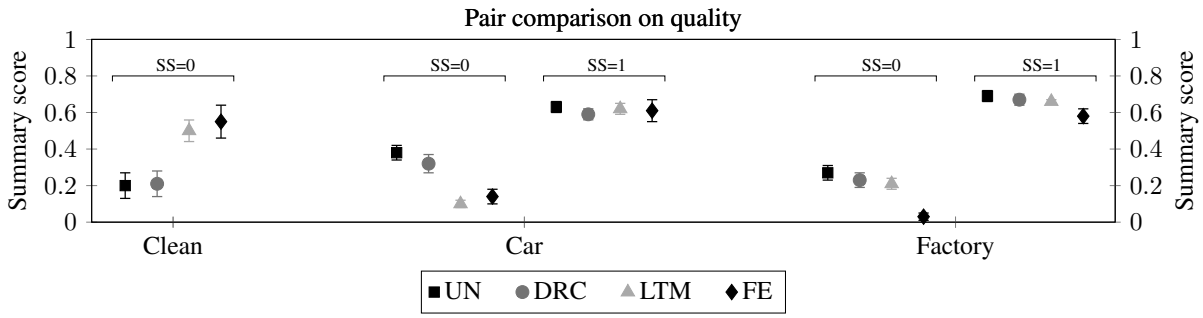

Figure 3: The means and the standard errors of the mean for the summary scores on quality in all test conditions. The post-processing methods under comparison were unprocessed speech (UN), dynamic range compression (DRC), Lombard-tilt modelling (LTM) and the formant equalizing postfilter (FE). Spectral subtraction was used as a pre-processing step in the cases marked with SS=1. The far-end conditions without noise and in car and factory noise are shown from left to right.

(UN, DRC, LTM, FE), the near-end noise level (moderate, severe) and the speaker gender (male, female) as fixed factors and the listener as a random factor. The summary scores for quality were analyzed using ANOVA with the method (UN, DRC, LTM, FE) as a fixed factor and the listener as a random factor. For both the WER and the pair comparison scores, an additional fixed factor in the analysis was the use of spectral subtraction (SS = 0, SS = 1) in the far-end. The data was checked for sphericity and the Greenhouse-Geisser correction was applied if needed. Post hocs were conducted using Tukey's test with 95% significance level. In the following, all the relevant results from the statistical analysis significant with 95% level are reported.

The intelligibility scores in the clean far-end condition were affected by the near-end noise level [$F(1, 10) = 106.25$], the method [$F(1.94, 19.50) = 18.58$], and the speaker [$F(1, 10) = 21.60$] as well as the interaction between the method and the speaker [$F(3, 30) = 3.61$]. Overall, FE and LTM received lower error rates than UN and DRC. On closer inspection, the difference between the methods was significant only with female speakers. In the far-end noise condition with car noise, the analysis showed that the near-end noise level [$F(1, 10) = 64.71$], the method [$F(3, 30) = 10.34$], the speaker [$F(1, 10) = 28.05$], and the use of spectral subtraction [$F(1, 10) = 112.71$] as well as the interaction between the near-end noise level and the use of spectral subtraction [$F(1, 10) = 10.56$] had a significant effect on intelligibility. Overall, FE and LTM received lower error rates than UN and DRC. The intelligibility scores were on average improved by the use of spectral subtraction. This trend was observed in both moderate and severe near-end noise. In the far-end noise condition with factory noise, the analysis showed that the near-end noise level [$F(1, 10) = 44.87$] and the speaker [$F(1, 10) = 34.03$] as well as the interactions between the near-end noise level and the speaker [$F(1, 10) = 16.88$] and between the speaker and the use of spectral subtraction [$F(1, 10) = 6.89$] had a significant effect on WER scores. All the findings are illustrated in Fig. 2.

The summary scores on quality, in the clean far-end noise condition, were affected by the method [$F(1.34, 13.40) = 4.89$]. Overall, FE and LTM were rated higher than DRC and UN. In the far-end condition with car noise, the quality scores were affected by the method [$F(1.63, 16.34) = 6.51$], the use of spectral subtraction [$F(1, 10) = 90.78$] as well as the interaction between the method and the use of spectral subtraction [$F(3, 30) = 9.78$]. Overall, LTM and FE were rated lower than UN and the samples with spectral subtraction received higher scores than the ones without it. Closer inspection of the interaction term revealed that both DRC and UN were rated higher in quality than LTM and FE when spectral subtraction was not used but this difference was not observed when spectral subtraction was used. In the far-end condition with factory noise, the summary score was affected by the method [$F(1.85, 18.46) = 11.98$], the use of spectral subtraction [$F(1, 10) = 226.01$] and by the interaction of the method and the use of spectral subtraction [$F(1.92, 19.15) = 3.69$]. Overall, FE was rated lower than the other methods and closer inspection of the interaction term showed that this difference was seen when spectral subtraction was not used. On average, conditions with spectral subtraction were rated higher than conditions without it. These observations are visualized in Fig. 3.

4. Discussion

The performance of three post-processing methods was compared to unprocessed speech in the presence of both near-end and far-end noise with and without spectral subtraction. The evaluation contained both a WER test and a pair comparison test on quality with different far-end noise conditions.

In the clean far-end condition, two post-processing methods, FE and LTM, performed better in terms of WER compared to UN and DRC. This difference was also observed with car noise in the far-end whereas with factory noise no significant differences between the methods were found. In car noise, the spectral subtraction improved the intelligibility scores but in factory noise, the SS has no significant impact, although it even tends to increase the error rates slightly in some cases. In the far-end condition with car noise, the improvement in WER with the spectral subtraction is relatively small in some conditions with FE and LTM, which were the most efficient post-processing techniques in clean far-end conditions. This suggests that spectral subtraction is not necessarily beneficial in all cases even in stationary far-end noise. However, the use of SS significantly improved the quality scores compared to the conditions without SS. This was especially notable for FE and LTM, which were rated lower than UN and DRC with SS = 0 in the noisy far-end conditions but received similar scores when spectral subtraction was used.

5. Acknowledgements

This work was supported by the Academy of Finland (projects 256961, 284671).

6. References

[1] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachtagung Sprachkommunikation*, 2010.

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

[3] 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 *Proceedings of Interspeech*, 2013, pp. 3577–3581.

[4] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proceedings of Interspeech*, 2012.

[5] J. Hall and J. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 280–285, 2010.

[6] 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," *Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3990–4001, 2012.

[7] W. V. Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

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

[9] E. Jokinen, M. Takanen, M. Vainio, and P. Alku, "An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech," *Computer, Speech and Language*, vol. 28, no. 2, pp. 619–628, 2014.

[10] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from Lombard and clear speaking styles," *Computer, Speech and Language*, vol. 28, no. 2, pp. 629–647, 2014.

[11] E. Jokinen, U. Remes, M. Takanen, K. Palomäki, M. Kurimo, and P. Alku, "Spectral tilt modelling with extrapolated GMMs for intelligibility enhancement of narrowband telephone speech," in *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 164–168.

[12] ——, "Spectral tilt modelling with GMMs for intelligibility enhancement of narrowband telephone speech," in *Proceedings of Interspeech*, 2014, pp. 2036–2040.

[13] A. Griffin, T.-C. Zorilă, and Y. Stylianou, "Improved face-to-face communication using noise reduction and speech intelligibility enhancement," in *Proceedings of IEEE International Conference on Acoustics, Speech and, Signal Processing (ICASSP)*, 2015, pp. 5103–5107.

[14] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7–8, pp. 588 – 601, 2007.

[15] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*. Springer–Verlag, 2005.

[16] D. Ellis. (2003) Dynamic time warp (DTW) in Matlab. Visited 16.03.2014. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/

[17] E. Jokinen, U. Remes, and P. Alku, "Comparison of gaussian process regression and gaussian mixture models in spectral tilt modelling for intelligibility enhancement of telephone speech," in *Proceedings of Interspeech*, 2015, pp. 85–89.

[18] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[19] *Adaptive multi-rate (AMR) speech codec; Transcoding functions*, 3rd Generation Partnership Project, Valbonne, France, 2008, version 8.0.0.

[20] B. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 1, pp. 22–32, 1969.

[21] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[22] M. Brookes. (1997) Voicebox: Speech processing toolbox for Matlab. Visited 16.01.2016. [Online]. Available: www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[23] A. Varga and H. 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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[24] *EVS Permanent Document EVS-8a:Test plans for qualification phase including host lab specification*, 3rd Generation Partnership Project - Meeting S4-72, Valencia, Spain, 2013, version 1.1.2. Visited: 12.02.2016. [Online]. Available: http://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/EVS_Permanent_Documents/.

[25] 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," *Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1742–1750, 2005.

[26] *EVS Permanent Document EVS-7a: Processing functions for qualification phase*, 3rd Generation Partnership Project - Meeting S4-72, Valencia, Spain, 2013, version 1.3. Visited: 12.02.2016. [Online]. Available: http://www.3gpp.org/ftp/tsg_sa/WG4_CODEC/EVS_Permanent_Documents/.

[27] *Recommendation G.191: Software tools for speech and audio coding standardization*, International Telecommunication Union, Geneva, Switzerland, September 2005.

[28] *Recommendation P.56: Objective measurement of active speech level*, International Telecommunication Union, Geneva, Switzerland, March 1993.

[29] *Specification TS 26.104: ANSI-C code for the floating-point Adaptive Multi-Rate (AMR) speech codec*, 3rd Generation Partnership Project, Valbonne, France, 2009, version 9.0.0.