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

Joint enhancement and coding of speech by incorporating Wiener filtering in a CELP codec

Johannes Fischer[*] *and Tom Bäckström*[*†]

[*]International Audio Laboratories Erlangen, Friedrich-Alexander-Universität (FAU)
[†]Fraunhofer IIS, Erlangen, Germany
first.lastname@audiolabs-erlangen.de

Abstract

The performance of speech communication applications in the field of mobile devices is often hampered by background noises and distortions. Therefore, noise attenuation methods are commonly used as a pre-processing method, cascaded with the speech-codec. We demonstrate that the performance of such combinations of speech enhancement and coding methods can be improved by joining the two methods into a single block. The proposed method is based on incorporating Wiener filtering into the objective function used for optimization of the quantization in code excited linear prediction (CELP)-based codecs. The benefits are that 1) the non-linear components of CELP codecs, including quantization and error feedback, are taken into account in the joint minimization function thereby improving quality and 2) by merging blocks both delay and computational complexity can be minimized. Our experiments demonstrate that the proposed joint enhancement and coding approach consistently improves subjective and objective quality. The proposed method is compatible with any CELP-based codecs without changing the bit-stream, whereby it can be readily applied in mobile phones or speech communication devices applying the concepts of CELP codecs for improving perceptual quality in adverse conditions.

Index Terms: speech coding, speech enhancement, code-excited linear prediction, Wiener filtering

1. Introduction

As speech communication devices have become ubiquitous, and are likely to be used in adverse conditions, the demand for speech enhancement methods has increased. Consequently, for example, in mobile phones it is by now common to use noise attenuation methods as a pre-processing step for all subsequent speech processing such as speech coding. There exist various approaches which incorporate speech enhancement into speech coders [1, 2, 3, 4]. While such designs improve transmitted speech quality, a joint minimization of quantization noise and interference has been difficult. The proposed method avoids accumulation of errors due to cascaded processing, as a joint minimization of interference and quantization distortion is realized by an optimal Wiener filtering in a perceptual domain.

The goal of speech codecs is to allow transmission of high quality speech with a minimum amount of transmitted data. To reach this goal we need efficient representations of the signal, such as modelling the spectral envelope of speech signals by linear prediction, the fundamental frequency by a long-time predictor and the remainder with a noise codebook. This representation is the basis of speech codecs using the CELP paradigm, which is used in major speech coding standards such as Adaptive Multi-Rate (AMR), AMR-Wide-Band (AMR-WB), Unified Speech and Audio Coding (USAC) and Enhanced Voice Service (EVS) [5, 6, 7, 8, 9, 10, 11].

For natural speech communication, speakers often use devices in hands-free modes. In such scenarios the microphone is usually far from the mouth, whereby the speech signal is easily distorted by interferences such as reverberation or background noise. The degradation does not only affect the perceived speech quality, but also the intelligibility of the speech signal and can therefore severely impede the naturalness of the conversation. To improve the communication experience, it is beneficial to apply speech enhancement methods to attenuate noise and reduce the effects of reverberation. The field of speech enhancement is mature and plenty of methods are readily available [12]. However, a majority of existing algorithms are based on transforms like the short-time Fourier transform (STFT), that apply overlap-add based windowing schemes, whereas CELP codecs model the signal with a linear predictive filter and apply windowing only on the residual. Such fundamental differences make it difficult to merge enhancement and coding methods. Yet it is clear that joint optimization of enhancement and coding can potentially improve quality, reduce delay and computational complexity.

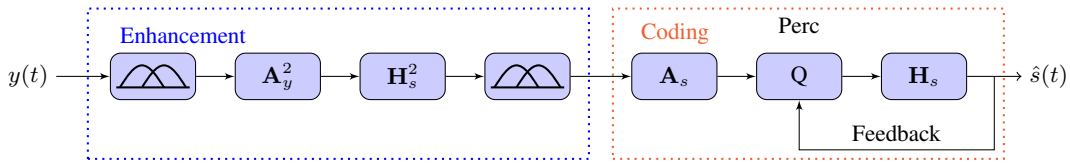
In this paper, we describe a method for joint enhancement and coding, based on Wiener filtering [12] and CELP coding. The advantages of this fusion are that 1) inclusion of Wiener filtering in the processing chain does not increase the low algorithmic delay of the CELP codec, and that 2) the joint optimization simultaneously minimizes distortion due to quantization and background noise. Moreover, the computational complexity of the joint scheme is lower than the one of the cascaded approach. The implementation relies on our recent work on residual-windowing in CELP-style codecs [13, 14, 15], which allows us to incorporate the Wiener filtering into the filters of the CELP codec in a new way. We demonstrate that both objective and subjective quality is improved in comparison to a cascaded system.

2. Code Excited Linear Prediction

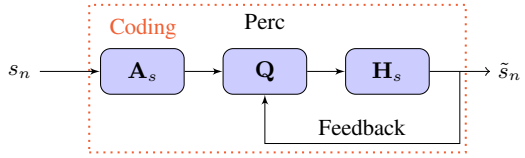
Speech codecs based on the CELP paradigm utilize a speech production model that assumes that the correlation, and therefore the spectral envelope of the input speech signal s_n can be modeled by a linear prediction filter with coefficients $\mathbf{a} = [\alpha_0, \alpha_1, \ldots, \alpha_M]^T$ where M is the model order, determined by the underlying tube model [16]. The residual $r_n = a_n * s_n$, the part of the speech signal that can not be predicted by the linear prediction filter, is then quantized using vector quantization.

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

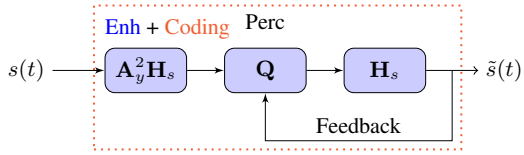
http://dx.doi.org/10.21437/Interspeech.2016-245



(a) Cascaded Enhancement and Coding.



(b) CELP Speech Coding



(c) Joint Enhancement and Coding

Figure 1: Illustrations of the cascaded and the joint enhancement/coding approaches. Here $\mathbf{A}_y$ and $\mathbf{A}_s$ represent the whitening filters of the noisy and clean signals, respectively, and $\mathbf{H}_y$ and $\mathbf{H}_s$ are the reconstruction filters, their corresponding inverses.

The linear predictive filter $\mathbf{a}_s$ for one frame of the input signal $\mathbf{s}$ can be obtained, minimizing

$$\min_{\mathbf{a}_s} \mathcal{E}\left\{\|\mathbf{s}^*\mathbf{a_s}\|^2 - 2\sigma_s^2(\mathbf{u}^*\mathbf{a_s} - 1)\right\}, \tag{1}$$

where $\mathbf{u} = \begin{bmatrix}1 & 0 & 0 & \dots & 0\end{bmatrix}^T$. The solution is

$$\mathbf{a}_s = \sigma_e^2 \mathbf{R}_{ss}^{-1}\mathbf{u}, \tag{2}$$

where $\mathbf{R}_{ss}$ is the autocorrelation matrix and σ_e^2 is the variance of the residual signal $\mathbf{e}_s$.

By defining the convolution matrix $\mathbf{A}_s$ using the coefficients the filter coefficients α of $\mathbf{a_s}$

$$\mathbf{A}_s = \begin{bmatrix} 1 & 0 & \dots & & 0 \\ \alpha_1 & \ddots & & & \vdots \\ \alpha_2 & \ddots & 1 & \ddots & \\ \vdots & \ddots & \alpha_1 & 1 & 0 \\ \alpha_M & \dots & \alpha_2 & \alpha_1 & 1 \end{bmatrix}, \tag{3}$$

the residual signal can be obtained by multiplying the input speech frame with the convolution matrix $\mathbf{A}_s$

$$\mathbf{e}_s = \mathbf{A}_s\mathbf{s}. \tag{4}$$

Windowing is here performed as in CELP-codecs by subtracting the zero-input response from the input signal and reintroducing it in the resynthesis [15].

The multiplication in Equation 4 is identical to the convolution of the input signal with the prediction filter, and therefore corresponds to FIR filtering. The original signal can be reconstructed from the residual, by a multiplication with the reconstruction filter $\mathbf{H}_s$

$$\mathbf{s} = \mathbf{H}_s\mathbf{e}_s, \tag{5}$$

where $\mathbf{H}_s$, consists of the impulse response $\eta = [1, \eta_1, \dots, \eta_{N-1}]$ of the prediction filter

$$\mathbf{H}_s = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \eta_1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \eta_{N-1} & \dots & \eta_1 & 1 \\ \vdots & & \vdots & \vdots \end{bmatrix} \tag{6}$$

such that this operation corresponds to IIR filtering.

For quantization of the residual vector, we use vector quantization. To determine that residual vector $\hat{\mathbf{e}}_s$ which minimizes perceptual distortion, we can optimize

$$\min_{\hat{\mathbf{e}}_s} \|\mathbf{WH}(\hat{\mathbf{e}}_s - \mathbf{e}_s)\|^2, \tag{7}$$

where $\mathbf{e}_s$ is the unquantized residual and $\mathbf{W}(z) = A(0.92z)$ is the perceptual weighting filter, as used in the AMR-WB speech codec [6].

3. Application of Wiener Filtering in a CELP codec

For the application of single-channel speech enhancement, we assume that the acquired microphone signal y_n, is an additive mixture of the desired clean speech signal s_n and some undesired interference v_n, such that $y_n = s_n + v_n$. In the Z-domain, we have equivalently $Y(z) = S(z) + V(z)$.

By applying a Wiener filter $B(z)$ we want to reconstruct the speech signal $S(z)$ from the noisy observation $Y(z)$ by filtering, such that the estimated speech signal is $\hat{S}(z) := B(z)Y(z) \approx S(z)$. The minimum mean square solution for the Wiener filter is [12]

$$B(z) = \frac{|S(z)|^2}{|S(z)|^2 + |V(z)|^2}, \tag{8}$$

given the assumption that the speech and noise signals s_n and v_n, respectively, are uncorrelated.

In a speech codec, we have an estimate of the power spectrum available of the noisy signal y_n, in the form of the impulse response of the linear predictive model $|A_y(z)|^{-2}$. In other words, $|S(z)|^2 + |V(z)|^2 \approx \gamma|A_y(z)|^{-2}$ where γ is a scaling coefficient. The noisy linear predictor can be calculated from the autocorrelation matrix $\mathbf{R}_{yy}$ of the noisy signal as usual.

We furthermore need to estimate the power spectrum of the clean speech signal $|S(z)|^2$ or equivalently, the autocorrelation matrix $\mathbf{R}_{ss}$ of the clean speech signal. Enhancement algorithms often assume that the noise signal is stationary, whereby we can estimate the autocorrelation of the noise signal as $\mathbf{R}_{vv}$, from a non-speech frame of the input signal. The autocorrelation matrix of the clean speech signal $\mathbf{R}_{ss}$ can then be estimated as $\hat{\mathbf{R}}_{ss} = \mathbf{R}_{yy} - \mathbf{R}_{vv}$. Here we need to make the usual precautions to ensure that $\hat{\mathbf{R}}_{ss}$ remains positive definite.

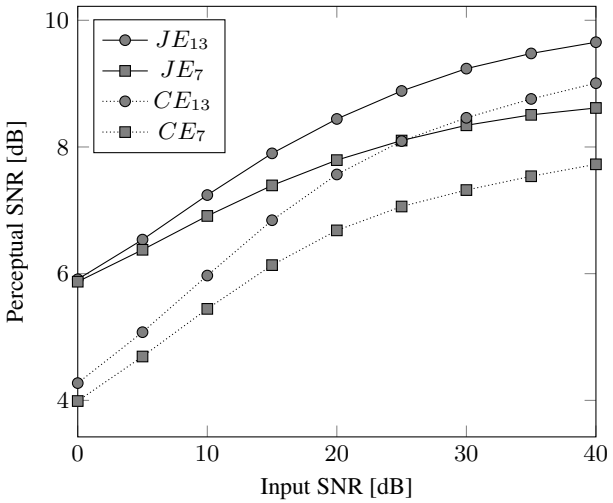


Figure 2: The perceptual magnitude SNR, as defined in Equation 12 for the proposed joint approach (JE) and the cascaded method (CE). The input signal was degraded by non-stationary car noise, and the results are presented for two different bitrates 7.2 kbit/s (7) and 13.2 kbit/s (13).

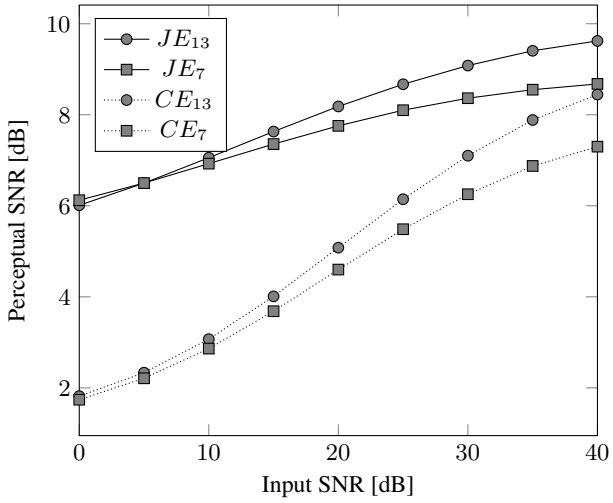


Figure 3: The perceptual magnitude SNR, as defined in Equation 12 for the proposed joint approach (JE) and the cascaded method (CE). The input signal was degraded by stationary white noise, and the results are presented for two different bitrates 7.2 kbit/s (7) and 13.2 kbit/s (13).

Using the estimated autocorrelation matrix for clean speech $\hat{\mathbf{R}}_{ss}$, we can then determine the corresponding linear predictor, which impulse response in Z-domain is $\hat{A}_s^{-1}(z)$. We thus have $|S(z)|^2 \approx |\hat{A}_s(z)|^{-2}$ and Eq. 8 can be written as

$$B(z) \approx \frac{|\hat{A}_s(z)|^{-2}}{|A_y(z)|^{-2}} = \frac{|A_y(z)|^2}{|\hat{A}_s(z)|^2}. \quad (9)$$

In other words, by filtering twice with the predictors of the noisy and clean signals, in FIR and IIR mode respectively, we obtain a Wiener estimate of the clean signal.

Let us denote the convolution matrices, corresponding to FIR filtering with predictors $\hat{A}_s(z)$ and $A_y(z)$ by $\mathbf{A}_s$ and $\mathbf{A}_y$, respectively. Similarly, let $\mathbf{H}_s$ and $\mathbf{H}_y$ be the respective convolution matrices corresponding to predictive filtering (IIR). Using these matrices, we can illustrate conventional CELP coding with a flow diagram as in Fig.1(b). Here we filter the input signal s_n with $\mathbf{A}_s$ to obtain the residual, quantize it and reconstruct the quantized signal by filtering with $\mathbf{H}_s$.

The conventional approach to combining enhancement with coding is illustrated in Fig. 1(a), where Wiener filtering is applied as a pre-processing block before coding.

Finally, in the proposed approach we combine Wiener filtering with CELP type speech codecs. Comparing the cascaded approach from Fig. 1(a) to the joint approach, illustrated in (b), it is evident that the additional overlap add windowing (OLA) windowing scheme can be omitted. Moreover, the input filter $\mathbf{A}_s$ at the encoder cancels out with $\mathbf{H}_s$. Therefore, as shown in Fig. 1(c), the estimated clean residual signal $\tilde{\mathbf{e}} = \mathbf{A}_{\mathrm{y}}^2\mathbf{H}_s\mathbf{y}$ follows by filtering the deteriorated input signal $\mathbf{y}$ with the filter combination $\mathbf{A}_{\mathrm{y}}^2\mathbf{H}_s$. Therefore, the error minimization follows:

$$\min_{\hat{\mathbf{e}}} \|\mathbf{W}\mathbf{H}_s(\hat{\mathbf{e}} - \tilde{\mathbf{e}})\|^2. \quad (10)$$

Thus, this approach jointly minimizes the distance between the clean estimate and the quantized signal, whereby a joint minimization of the interference and the quantization noise in the perceptual domain is feasible.

4. Experiments

The performance of the joint speech coding and enhancement approach was evaluated using both objective and subjective measures. In order to isolate the performance of the new method, we used a simplified CELP codec, where only the residual signal was quantized, but the delay and gain of the long term prediction (LTP), the linear predictive coding (LPC) and the gain factors were not quantized. The residual was quantized using a pair-wise iterative method, where two pulses are added consecutively by trying them on every position, as described in [17]. Moreover, to avoid any influence of estimation algorithms, the correlation matrix of the clean speech signal $\mathbf{R}_{ss}$ was assumed to be known in all simulated scenarios. With the assumption that the speech and the noise signal are uncorrelated, it holds that $\mathbf{R}_{ss} = \mathbf{R}_{yy} - \mathbf{R}_{vv}$. In any practical application the noise correlation matrix $\mathbf{R}_{vv}$ or alternatively the clean speech correlation matrix $\mathbf{R}_{ss}$ has to be estimated from the acquired microphone signal. A common approach is to estimate the noise correlation matrix in speech brakes, assuming that the interference is stationary.

The evaluated scenario consisted of a mixture of the desired clean speech signal and additive interference. We considered two types of interferences: stationary white noise and a segment of a recording of car noise from the Civilisation Soundscapes Library [18]. Vector quantization of the residual was performed with a bit-rate of 2.8 kbit/s and 7.2 kbit/s, corresponding to an overall bit-rate of 7.2 kbit/s and 13.2 kbit/s respectively for an AMR-WB codec [6]. A sampling-rate of 12.8 kHz was used for all simulations.

The enhanced and coded signals were evaluated using both objective and subjective measures, therefore a listening test was conducted and a perceptual magnitude signal-to-noise ratio (SNR) was calculated, as defined in Equation 12 and Equation 11. We used this perceptual magnitude SNR as the joint enhancement process has no influence on the phase of the filters, as both the synthesis and the reconstruction filters are

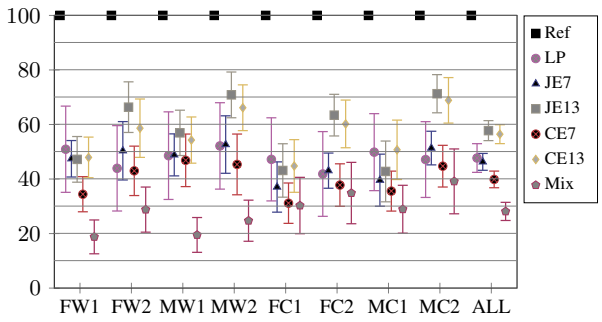


Figure 4: Illustration of the MUSHRA scores for the different English speakers (Female (F) and Male (M)), for two different interferences (White noise (W) and Car noise (C)), for two different input SNRs (10 dB (1) and 20 dB (2)). All items were encoded at two bit-rates 7.2 kbit/s (7) and 13.2 kbit/s (13), for the proposed joint approach (JE) and the cascaded enhancement (CE). Ref was the hidden reference, LP the 3.5 kHz low-pass anchor and Mix the distorted mixture.

bound to the constraint of minimum phase filters, as per design of prediction filters.

With the definition of the Fourier transform as operator $\mathcal{F}(\cdot)$, the absolute spectral values of the reconstructed clean reference and the estimated clean signal in the perceptual domain are:

$$S = |\mathcal{F}(\mathbf{W}\mathbf{H}_s\mathbf{e}_k)| \qquad \text{and} \qquad \hat{S} = |\mathcal{F}(\mathbf{W}\mathbf{H}_s\hat{\mathbf{e}}_k)|. \quad (11)$$

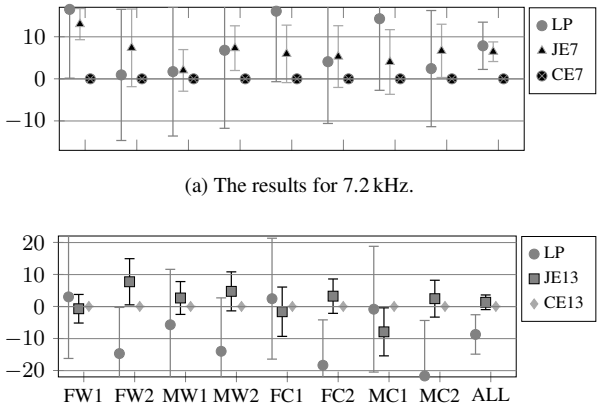
The definition of the modified perceptual signal to noise ratio (PSNR) is:

$$\mathrm{PSNR}_{\mathrm{ABS}} = 10\log_{10}\frac{\|S\|^2}{\|\hat{S}-S\|^2}. \quad (12)$$

For the subjective evaluation we used speech items from the test set used for the standardization of USAC [8], corrupted by white- and car-noise, as described above. We conducted a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [19] listening test with 14 participants, using STAX electrostatic headphones in a soundproof environment. The results of the listening test are illustrated in Figure 4 and the differential MUSHRA scores in Figure 5, showing the mean and 95% confidence intervals.

The absolute MUSHRA test results in Figure 4 show that the hidden reference was always correctly assigned to 100 points. The original noisy mixture received the lowest mean score for every item, indicating that all enhancement methods improved the perceptual quality. The mean scores for the lower bit-rate show a statistically significant improvement of 6.4 MUSHRA points for the average over all items in comparison to the cascaded approach. For the higher bit-rate, the average over all items shows an improvement, which however is not statistically significant.

To obtain a more detailed comparison of the joint and the pre-enhanced methods we present the differential MUSHRA scores in Figure 5, where the difference between the pre-enhanced and the joint methods is calculated for each listener and item. The differential results verify the absolute MUSHRA scores, by showing a statistically significant improvement for the lower bit-rate, whereas the improvement for the higher bit-rate is not statistically significant.



(a) The results for 7.2 kHz.

(b) The results for 13.2 kHz.

Figure 5: Differential MUSHRA scores, simulated over two different bit-rates, comparing the new joint enhancement (JE), to a cascaded approach (CE).

5. Conclusions

In this paper we have presented a method for joint speech enhancement and coding, which allows minimization of overall interference and quantization noise. In contrast, conventional approaches apply enhancement and coding in cascaded processing steps. Joining both processing steps is also attractive in terms of computational complexity, since repeated windowing and filtering operations can be omitted.

CELP type speech codecs are designed to offer a very low delay and therefore avoid an overlap of processing windows to future processing windows. In contrast, conventional enhancement methods, applied in the frequency domain rely on overlap-add windowing, which introduces an additional delay corresponding to the overlap length. The joint approach does not require overlap-add windowing, but uses the windowing scheme as applied in speech codecs [15], whereby we avoid the increase in algorithmic delay.

A known issue with the proposed method is that, in difference to conventional spectral Wiener filtering where the signal phase is left intact, the proposed method applies time-domain filters, which do modify the phase. Such phase-modifications can be readily treated by application of suitable all-pass filters. However, since we have not noticed any perceptual degradation attributed to phase-modifications, we have chosen to omit such all-pass filters to keep computational complexity low. Note, however, that in the objective evaluation, we measured perceptual magnitude SNR, to allow fair comparison of methods. This objective measure shows that the proposed method is on average three dB better than cascaded processing.

The performance advantage of the proposed method was further confirmed by the results of a MUSHRA listening test, which show an average improvement of 6.4 points. These results demonstrate that application of joint enhancement and coding is beneficial for the overall system in terms of both quality and computational complexity, while maintaining the low algorithmic delay of CELP speech codecs.

6. References

[1] M. Jeub and P. Vary, "Enhancement of reverberant speech using the CELP postfilter," in *Proc. ICASSP*, April 2009, pp. 3993–3996.

[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. ICASSP*, March 2012, pp. 1693–1696.

[3] R. Martin, I. Wittke, and P. Jax, "Optimized estimation of spectral parameters for the coding of noisy speech," in *Proc. ICASSP*, vol. 3, 2000, pp. 1479–1482 vol.3.

[4] H. Taddei, C. Beaugeant, and M. de Meuleneire, "Noise reduction on speech codec parameters," in *Proc. ICASSP*, vol. 1, May 2004, pp. I–497–500 vol.1.

[5] 3GPP, "Mandatory speech CODEC speech processing functions; AMR speech Codec; General description," 3rd Generation Partnership Project (3GPP), TS 26.071, 12 2009. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/26071.htm

[6] ——, "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions," 3rd Generation Partnership Project (3GPP), TS 26.190, 12 2009. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/26190.htm

[7] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov 2002.

[8] ISO/IEC 23003–3:2012, "MPEG-D (MPEG audio technologies), Part 3: Unified speech and audio coding," 2012.

[9] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefebvre, and B. Grill, "Unified speech and audio coding scheme for high quality at low bitrates," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 1–4.

[10] 3GPP, "TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12)," 3rd Generation Partnership Project (3GPP), TS 26.445, 12 2014. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/26445.htm

[11] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache, Y. Kamamoto, K. Kikuiri, S. Ragot, J. Faure, H. Ehara, V. Rajendran, V. Atti, H. Sung, E. Oh, H. Yuan, and C. Zhu, "Overview of the EVS codec architecture," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 5698–5702.

[12] J. Benesty, M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Springer, 2008.

[13] T. Bäckström, "Computationally efficient objective function for algebraic codebook optimization in ACELP," in *Proc. Interspeech*, Aug. 2013.

[14] ——, "Comparison of windowing in speech and audio coding," in *Proc. WASPAA*, New Paltz, USA, Oct. 2013.

[15] J. Fischer and T. Bäckström, "Comparison of windowing schemes for speech coding," in *Proc EUSIPCO*, 2015.

[16] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. ICASSP*. IEEE, 1985, pp. 937–940.

[17] T. Bäckström and C. R. Helmrich, "Decorrelated innovative codebooks for ACELP using factorization of autocorrelation matrix," in *Proc. Interspeech*, 2014, pp. 2794–2798.

[18] soundeffects.ch, "Civilisation soundscapes library," accessed: 23.09.2015. [Online]. Available: https://www.soundeffects.ch/de/geraeusch-archive/soundeffects.ch-produkte/civilisat

[19] *Method for the subjective assessment of intermediate quality levels of coding systems*, ITU-R Recommendation BS.1534, 2003. [Online]. Available: http://www.itu.int/rec/R-REC-BS.1534/en