Wideband Coding of Speech and Audio Signals using Bandwidth Extension Techniques

Thomas ESCH

Institute of Communication Systems and Data Processing, RWTH Aachen University,
Templergraben 55, 52056 Aachen, Germany

E-Mail: esch@ind.rwth-aachen.de

Abstract. *This paper presents a method for wideband (WB, acoustic bandwidth 50-7000 Hz) coding of speech and audio signals using techniques of bandwidth extension (BWE) with side information. On top of an already existing narrowband (NB, acoustic bandwidth 50-3400 Hz) speech or audio codec, additional information on the extension band (EB, acoustic bandwidth 3400-7000 Hz) is transmitted. Using this side information and the NB signal, the receiver is able to perform an estimation of the EB and therewith of the WB signal. Experimental results show that the proposed wideband codec greatly improves the subjectively perceived speech quality compared to the NB signal by increasing the total bit rate only marginally.*

Keywords

Wideband Speech and Audio Coding, Bandwidth Extension, Side Information

1. Introduction

The bandwidth of the transmitted audio signal in current public telephone systems is still limited to a frequency range of up to 3.4 kHz. This bandwidth limitation causes the characteristic sound of ”telephone speech” that reduces the speech intelligibility and is perceived as muffled compared to the original speech signal.
Listening experiments [1] have shown that increasing the bandwidth of the narrowband (NB) telephone signal greatly improves the subjectively perceived speech quality. This fact is utilized in [2], where the author presents an artificial bandwidth extension (BWE) system. This system extends the bandwidth from 3.4 kHz up to 7 kHz by estimating the extension band (EB) from parameters extracted from the NB signal and therewith exploits the high correlation between NB and EB. Results of this artificial BWE have shown that the speech quality could indeed be enhanced but now suffers from unnatural artifacts due to estimation errors of the EB, especially for fricative sounds.
Hence the wideband (WB) codec proposed in this paper extends the approach of [2] by estimating the EB in the receiver with additional information. Therefor the transmitter extracts a comparably low amount of side information from the EB, which is used in the receiver together with the NB signal to perform a more precise estimation of the EB.
The remainder of this paper is organized as follows: Section 2 introduces the principal procedure of BWE with side information, Section 3 describes the quantization and coding of the extracted side information and in Section 4 experimental results are presented.

2. BWE with Side Information

In this section, the main principle of the BWE with side information is described. It is assumed that the wideband speech signal s_{wb} (sampling frequency $f_s = 16$ kHz) is processed on a frame-by-frame basis with a frame size of 20 ms. For the computations in the frequency domain, the respective frame is first windowed with a sliding Hann-window (50% overlap) and then transformed into the frequency domain using zero-padding and an FFT of frame length $L = 1024$. After processing, the frame is transformed back into the time domain using an IFFT and the overlap-add method.
In the following the computation of the side information at the transmitter and the estimation of the EB signal at the receiver are descibed.

2.1. Parameter Extraction at the Transmitter

Before the side information can be extracted at the transmitter, the wideband signal s_{wb} is decomposed into a NB signal s_{nb} and an EB signal s_{eb} (e.g. by low-pass and band-pass filtering). For the coding of the NB signal, an already existing narrowband speech or audio codec (e.g., G.711 [3]) can be used. The extraction of the side information at the transmitter is depicted in Fig. 1. From the quantized NB signal $\hat{s}_{nb}$ the pitch frequency f_P is estimated and a voiced/unvoiced classification $S \in \{0, 1\}$ is performed for each frame. Well-known techniques from the literature can be used for this (see, e.g., [4],[5]).
Voiced segments of speech contain sinusoids at multiples of the pitch frequency f_P up to a certain frequency f_c. At higher frequencies, the spectrum is more flat and noise like. Fig. 2 depicts an example of the magnitude of the short-term spectrum of a WB signal. It can be seen that the spectrum is harmonic up to a cut-off frequency of $f_c \approx 5.4$ kHz.

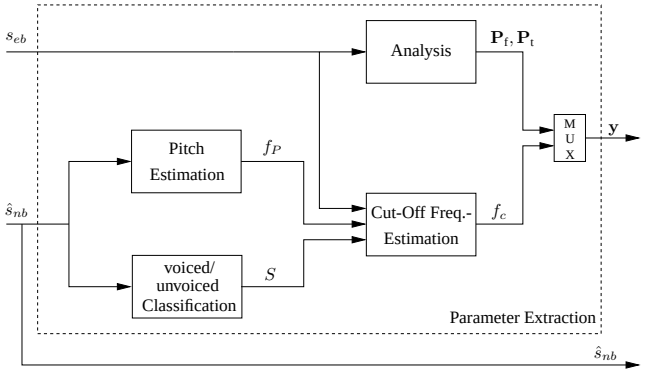

Fig. 1. Parameter Extraction at the transmitter.

The cut-off frequency needs to be calculated only for voiced speech segments. In unvoiced segments f_c is set to 0 Hz. To compute f_c, frequency bands with a bandwidth of 1 kHz are extracted from the short-term spectrum of s_{eb} using a sliding Hann-window. Afterwards the cepstrum [6] is calculated for each frequency band. If the frequency band contains a harmonic structure, the cepstrum shows a peak at the position of the pitch period $T_p = \frac{1}{f_P}$. Therefore the cepstrum at position T_p is compared to an empirically determined threshold in order to compute the cut-off frequency.

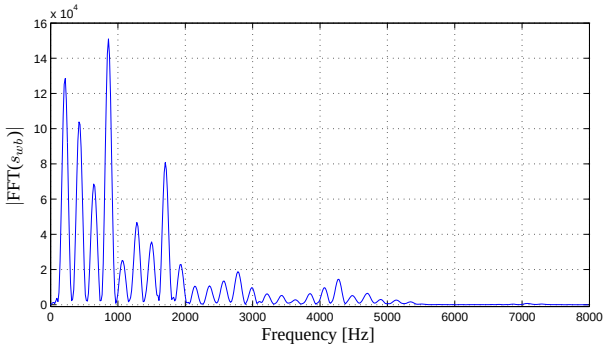

Fig. 2. Magnitude of the short-term spectrum of s_{wb}

In addition to the cut-off frequency, the side information also contains parameters describing the time and frequency envelope ($\mathbf{P}_\mathrm{t}$ and $\mathbf{P}_\mathrm{f}$) of the EB signal. In Fig. 1 this parameter extraction is performed by the Analysis block. The frequency envelope can for example be calculated by computing the signal power in sub-bands of the EB frequency range whereas the sub-bands are generated by applying a sliding window in the frequency domain. In a similar manner the time envelope may be determined by calculating the power of short-term windowed segments in the time domain. A more detailed description of how to extract the envelopes can be found e.g. in [7]. Due to the fact that both time and frequency envelopes are computed, the resolution of the extracted envelopes can be adjusted arbitrarily.
The parameters $\mathbf{P}_\mathrm{t}$, $\mathbf{P}_\mathrm{f}$ and f_c form a 115-dimensional parameter vector $\mathbf{y}$ every frame. Using the techniques described in Section 3 this parameter vector $\mathbf{y}$ is quantized, coded and transmitted.

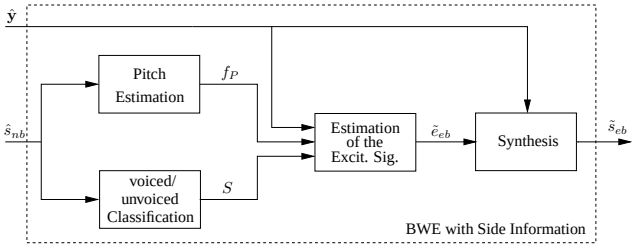

Fig. 3. Block diagram of BWE with side information at the receiver.

2.2. Estimation of the EB at the Receiver

The bandwidth extension at the receiver with the extracted side information is shown in Fig. 3. Assumed that no transmission error has occured, the quantized parameter vector $\hat{\mathbf{y}}$ and the quantized NB signal $\hat{s}_{nb}$ are available. In an analogous manner to the transmission side the pitch frequency f_P is estimated and a voiced/unvoiced classification S is performed from $\hat{s}_{nb}$ for every frame.
According to the simplifying linear model of speech production [6], the excitation signal $\tilde{e}_{eb}$ of the EB is estimated, which is depicted more detailed in Fig. 4. Depending on the classification S, the excitation signal is either voiced or unvoiced. For producing unvoiced sounds, a white noise excitation signal n is used. If S classifies a speech segment as voiced the excitation signal $t' + n'$ consists of a periodic signal t up to the cut-off frequency f_c and of a noise like signal n at higher frequencies. The periodicity of t is thereby described by the pitch period T_P. In order to get an estimation of the EB excitation signal $\tilde{e}_{eb}$ the generated WB excitation signal $\tilde{e}_{wb}$ is band-pass filtered.

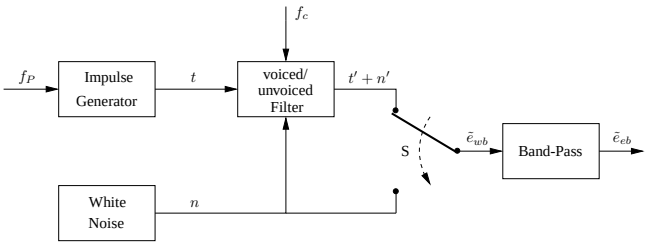

Fig. 4. Estimation of EB excitation signal.

After the excitation signal $\tilde{e}_{eb}$ has been generated, time and frequency envelope shaping is performed by a synthesis filter, see Fig. 3 and Fig. 5 for more details. Therefore the parameter vectors $\mathbf{P}_\mathrm{t}$ and $\mathbf{P}_\mathrm{f}$ are extracted from the side information. More information about the envelope shaping can be found in [7].

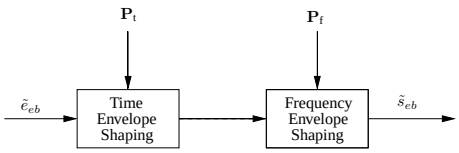

Fig. 5. Time and Frequency Envelope shaping.

Finally, the estimation of the wideband speech signal $\tilde{s}_{wb}$ is performed by adding the quantized narrowband signal $\hat{s}_{nb}$ and the estimated extension band signal $\tilde{s}_{eb}$ after delay compensation.
Experimental results have shown that without quantizing and coding of the NB signal and the parameter vector $\mathbf{y}$, there is almost no degradation of speech quality with the proposed method compared to the original speech signal.

3. Quantization and Coding

Due to the fact that the parameters in $\mathbf{y}$ are correlated with the narrowband signal s_{nb}, data rate would be wasted if the coding of $\mathbf{y}$ would be performed independently of s_{nb}. As depicted in Fig. 6 the side information $\mathbf{y}$ is therefore quantized and coded with the aid of a feature vector $\mathbf{x}$ that is extracted from the quantized narrowband signal at both the transmitter and the receiver side.

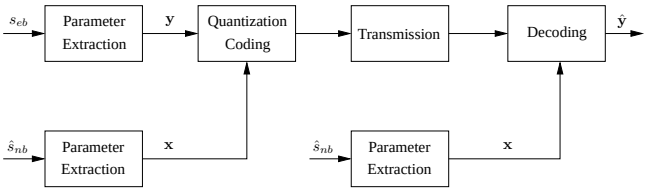

Fig. 6. Quantization and Coding of $\mathbf{y}$ with the aid of the feature vector $\mathbf{x}$.

The vector $\mathbf{x}$ contains almost the same types of parameters as $\mathbf{y}$, namely the time and frequency envelope of the quantized narrowband signal, calculated in an analogous manner as above. With the proposed coding scheme the mutual information $I(\mathbf{x};\mathbf{y})$ can be extracted from $\hat{s}_{eb}$ and only the amount of information described by the conditional entropy $H(\mathbf{y}|\mathbf{x}) = H(\mathbf{y}) - I(\mathbf{x};\mathbf{y})$ has to be transmitted.

The utilized quantization and coding scheme at the transmitter is depicted in Fig. 7 and in Fig. 8 the decoding at the receiver is shown. Thereby the parameter vector $\mathbf{y}$ is estimated from the feature vector $\mathbf{x}$ by linear mapping [8]. With linear mapping an estimate $\tilde{\mathbf{y}}$ of the vector $\mathbf{y}$ can be derived from the vector $\mathbf{x}$ by the transformation

$$\tilde{\mathbf{y}} = \mathbf{H}^T \cdot \mathbf{x}. \tag{1}$$

The dimension of the transformation matrix $\mathbf{H}$ is $b \times d$ with b the dimension of $\mathbf{x}$ and d the dimension of $\mathbf{y}$. The matrix $\mathbf{H}$ is derived during offline-training and contains the a priori knowledge of the dependencies between $\mathbf{x}$ and $\mathbf{y}$ [8].
As can be seen in Fig. 7 the estimation $\tilde{\mathbf{y}}$ is subtracted from

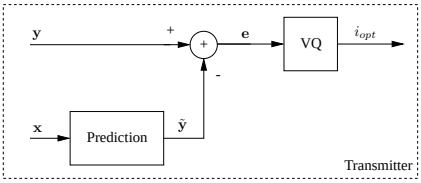

Fig. 7. Predictive Coding and Quantization at the transmitter.

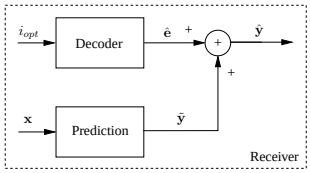

Fig. 8. Predictive Decoding at the receiver.

the original parameter vector $\mathbf{y}$ and only the resulting error vector $\mathbf{e}$ is vector quantized. The gain, which is achieved due to the fact that the variance of $\mathbf{e}$ is much lower than the variance of $\mathbf{y}$ can be used, e.g., to reduce the bit rate.
After decoding the optimal index i_{opt} from the vector quantizer (VQ) by table lookup, the receiver adds the quantized error vector $\hat{\mathbf{e}}$ and the estimated parameter vector $\tilde{\mathbf{y}}$ and gets the quantized parameter vector $\hat{\mathbf{y}}$ therewith (see Fig. 8).

To reduce the computational load of the system due to the high dimension of the vector $\mathbf{y}$ a linear discriminant analysis (LDA) can be used before vector quantization in order to reduce the dimension of $\mathbf{y}$.

4. Experimental Results

In two informal listening tests the speech quality of the proposed wideband codec was analyzed and compared to other narrowband and wideband speech codecs. To do this, the Multi Stimulus Test with Hidden Reference and Anchors (MUSHRA [9]) was used in two different scenarios. In each scenario twelve assessors were asked to detect any percetible annoyance of artifacts which may be present in the signal compared to the original (reference) signal. The listeners were required to score every speech signal with a number in the range of 0 (bad speech quality) to 100 (excellent speech quality). Each scenario included 5 different codecs that had to be analyzed while for each codec 5 different speech samples from different speakers and in different languages were available.

The training of the VQ and of the transformation matrix $\mathbf{H}$ from Section 3 was performed with samples from the NTT-AT data base [10]. Approximately 30 minutes of speech from different speakers and in different languages were taken therefrom. The training data did not include the speech samples used for the listening tests. In each scenario it was assumed that no transmission error occured.

The configuration of the two scenarios was as follows:

Scenario 1

- Original WB signal (Ref)
- G.722 [11] (G722 64)
  WB codec (bit rate $R_T = 64$ kbit/s)
- G.711 (G711 64)
  NB codec (bit rate $R_T = 64$ kbit/s)
- G.711 + BWE (G711+BWE 64.25)
  The NB codec G.711 (bit rate R =64 kbit/s) was used to transmit the NB signal and the side information for

BWE was transmitted with R =250 bit/s, i.e., the total bit rate was R_T =64.25 kbit/s.

- unquantized NB + BWE (uNB+BWE)
  The NB was transmitted without quantization and coding, the side information for BWE was transmitted with R =250 bit/s.

Scenario 2

- Original WB signal (Ref)
- AMR Wideband [12] (AMR-WB 14.25)
  WB codec (bit rate $R_T = 14.25$ kbit/s)
- GSM Enhanced Full Rate [13] (EFR 12.2)
  NB codec (bit rate $R_T = 12.2$ kbit/s)
- Enhanced Full Rate + BWE (EFR+BWE 14.2)
  The NB codec EFR (bit rate R =12.2 kbit/s) was used to transmit the NB signal and the side information for BWE was transmitted with R =2 kbit/s, i.e., the total bit rate was R_T =14.2 kbit/s.
- unquantized NB + BWE (uNB+BWE)
  The NB was transmitted without quantization and coding, the side information for BWE was transmitted with R =2 kbit/s.

The results of the informal listening tests can be seen in Fig. 9 for scenario 1 and in Fig. 10 for scenario 2. In both figures the overall mean score and the 95% confidence intervall for each analyzed codec is plotted.
Fig. 9 shows that a comparably low bit rate for the side information and the proposed BWE is sufficient to improve the speech quality of the NB codec G.711 significantly. However, the amount of side information is too low to achieve results comparable to the WB codec G.722.
In scenario 2 the EFR codec was used to transmit the NB and a BWE was performed with $R = 2$ kbit/s for the side information. Again the BWE greatly improves the speech quality of the NB codec. Actually the proposed WB codec achieves results comparable to the AMR WB codec.

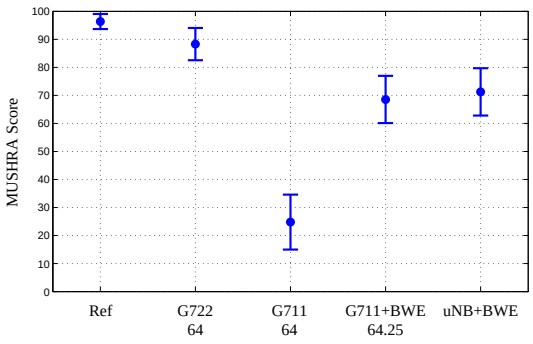

Fig. 9. MUSHRA results for Scenario 1.

5. Conclusion

In this paper a wideband codec for speech and audio signals was presented which extends an already existing narrowband codec. The narrowband codec is used for the transmission of the NB signal and a BWE at the receiver with additional side information from the transmitter is performed.

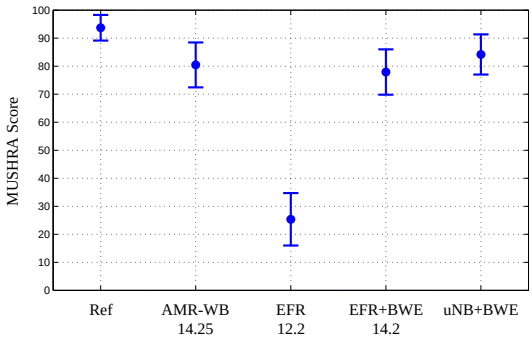

Fig. 10. MUSHRA results for Scenario 2.

Thereby the side information contains parameters describing the cut-off frequency and the time and frequency envelopes of the extension band signal.
Informal listening tests have shown that by increasing the total bit rate of a NB codec only marginally the subjectively perceived speech quality can be improved significantly and is absolutely comparable to other existing WB codecs. The advantage over the other tested WB codecs is the fact that embedded coding is possible, i.e. without transmission of the side information, the NB signal is still available at the receiver.

Acknowledgements

The author would like to thank the head of the Institute of Communication Systems and Data Processing, Prof. P. Vary. The underlying work of this paper is the result of a diploma thesis which was carried out at his institute.

References

[1] KREBBER, W. Sprachübertragungsqualität von Fernsprech-Handapparaten. *Ph.D. Thesis*, Aachen University (RWTH), 1995.

[2] JAX, P. Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds. *Ph.D. Thesis*, Aachen University (RWTH), 2002.

[3] ITU-T Recommendation G.711. Pulse code modulation (PCM) of voice frequencies, 1972.

[4] PAULUS, J. Codierung breitbandiger Sprachsignale bei niedriger Datenrate. *Ph.D. Thesis*, Aachen University (RWTH), 1997.

[5] HOELPER, C., FRANKORT, A., ERDMANN, C. Voiced/Unvoiced/Silence Classification for Offline Speech Coding, *Poster 2003*, Prague, 2003.

[6] VARY, P., MARTIN, R. *Digital Speech Transmission*. Wiley, 2006.

[7] JAX, P., GEISER, B., SCHANDL, S., TADDEI, H., VARY, P. An embedded scalable wideband codec based on the GSM EFR codec. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006.

[8] JAX, P. *Bandwidth extension for speech*. In LARSEN, E., AARTS, R. M. *Audio Bandwidth Extension*. Wiley, chapter 6, pp. 171-236, 2004.

[9] ITU-T Recommendation BS.1534-1. Method for the subjective assessment of intermediate quality level of coding systems, 2003.

[10] NTT Advanced Technology Corporation. Multi-lingual speech database for telephonometry. *Online at http://www.ntt-at.com/products_e/speech*, 1994.

[11] ITU-T Recommendation G.722. 7 khz audio coding within 64 kbit/s, 1988.

[12] 3GPP TS 26.171. AMR wideband speech codec; general description, 2001.

[13] ETSI Recommendation GSM 06.60. Enhanced full rate (EFR) speech transcoding, version 8.0.1, 1999.