A Scalable Wideband "Add-On" for the G.729 Speech Codec

Peter Jax[†]*, Bernd Geiser, Stefan Schandl*[1]*, Hervé Taddei*[2]*, and Peter Vary*

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{jax|geiser|vary}@ind.rwth-aachen.de

[1]Siemens AG, Vienna, Austria
stefan.schandl@siemens.com

[2]Siemens AG, Munich, Germany
herve.taddei@siemens.com

Abstract

We present a technique to enhance the perceived speech quality in *narrowband* (NB, cutoff frequency $f_c = 3.4\,\text{kHz}$) speech communication systems, e.g., implementing the ITU-T G.729A codec [1]. Therefore, *wideband* (WB, $f_c = 7\,\text{kHz}$) capability is introduced. The limited acoustic bandwidth of NB speech coding is extended using a fairly coarse description of the temporal and spectral envelopes of the missing high frequency band (3.4–7 kHz). These parameters are quantized, transmitted and then used at the receiver side to synthesize the high frequency components based on a synthetic "excitation signal". The parameter quantization is done using split Vector Quantization (VQ) in a transformed domain. The quantization scheme can be tailored to match the desired bitrate. We have implemented an example configuration and performed a listening test to compare the approach with common WB speech codecs.

1. Introduction

The constant increase of available data rates in telecommunication systems (e.g., in UMTS) allows to enhance the user experience with for example the transmission of a wideband signal instead of a narrowband version. This larger frequency band can be transmitted by designing a completely new wideband codec such as the *adaptive multirate wideband* (AMR-WB) codec [2]. Nevertheless, this option comes with a cost: Network equipment needs to be changed and/or transcoding has to be used. A different option is to take already deployed narrowband codecs (e.g., the ITU-T G.729A) into account and to build an *embedded wideband codec* on top of it. This is what ITU-T is currently promoting with the standardization work named G.729EV.

In order to encode the additional bandwidth, common wideband codecs often derive the highest frequency subband from lower frequency components (*bandwidth extension* (BWE), e.g. [3], [4]) and from a rather low amount of additional side information. Examples for speech and audio coding with BWE techniques are the Extended AMR-WB+ [5] and Enhanced aacPlus [6] codecs. Several related approaches exist, e.g., [7].

In this paper we also follow these paradigms. The *NB* frequency components (50–3400 Hz) are encoded by an enhanced version of the ITU-T G.729A [1] codec. This so-called G.729A+ codec has a gross bitrate of $8 + 4\,\text{kbit/s} = 12\,\text{kbit/s}$. The additional bitrate of 4 kbit/s is used by a *second fixed codebook search*. The *high* frequency band (3.4–7 kHz) is synthesized in the receiver using a BWE scheme, which is supported by a certain amount of side information.

Our wideband "add-on" extends the acoustic bandwidth of the narrowband output signal of the G.729A+ codec using only a coarse description of the *extension band* (EB) components (3.4–7 kHz). This description comprises time and frequency envelopes extracted from the EB components of the *original* wideband

[†]now with Thomson Corporate Research, Hannover

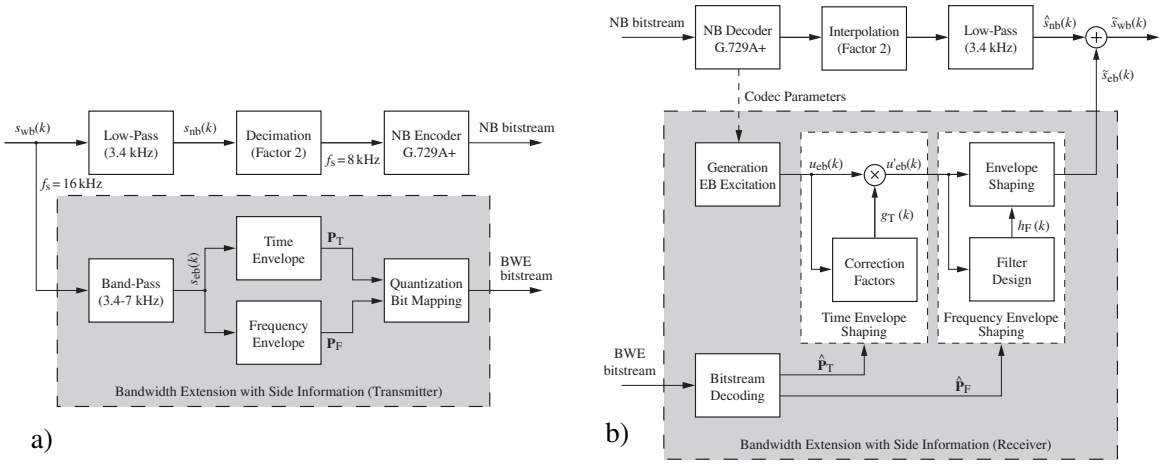

Fig. 1: BWE transmitter (a) and BWE receiver (b) with time and frequency envelope shaping of an artificially generated excitation signal.

speech signal every 20 ms. The extraction and the quantization of these parameters are described in Sec. 2. In the receiver, the EB signal components are synthesized by shaping the time and frequency envelopes of an artificially generated excitation signal, see Sec. 3. Here we have taken special care to produce smooth transitions in time and frequency domain.

2. BWE Transmitter

Fig. 1a) depicts the encoder side of the wideband "add-on". The wideband input signal $s_{\mathrm{wb}}(k)$ is sampled at 16 kHz. The upper signal path includes the G.729A+ narrow band encoding which operates at 8 kHz.

The lower signal path comprises the parameter extraction for both the time and the frequency envelopes of the EB signal components $s_{\mathrm{eb}}(k)$. First, the wideband input signal $s_{\mathrm{wb}}(k)$ is bandpass filtered (3.4–7 kHz) to isolate $s_{\mathrm{eb}}(k)$. Then, the time envelope of $s_{\mathrm{eb}}(k)$ is extracted in terms of *subframe energies* T_n (for each segment n of duration 2 ms). The frequency envelope is represented by *FFT subband energies* $F_{m,\nu}$ (for 10 ms segments with index m and for each of the $N_{\mathrm{sb}} = 10$ equally spaced sub-bands with index ν and with 375 Hz bandwidth, beginning at 3.5 kHz).

This envelope extraction yields a strongly correlated parameter vector for each speech frame (20 ms) which is transformed by a (long-term) *Principal Component Analysis* (PCA) matrix to obtain a rotated parameter space with mutually uncorrelated components. These are sorted by decreasing variance and split VQ with an optimized bit allocation is applied. This coding scheme is discussed in more detail in [8].

3. BWE Receiver

The BWE receiver is shown in Fig. 1b). The upper signal path includes the narrowband G.729A+ decoder whose output is upsampled to 16 kHz and lowpass filtered. The lower signal path contains the synthesis of the EB signal components. First, this procedure generates an excitation signal $u_{\mathrm{eb}}(k)$. The decoded BWE parameters $\hat{\mathbf{P}}_{\mathrm{T}}$ and $\hat{\mathbf{P}}_{\mathrm{F}}$ are then used to shape the time and frequency envelopes of $u_{\mathrm{eb}}(k)$ according to the properties of the original EB components. These blocks are subsequently described in the following sections.

3.1 Excitation Signal Generation

The excitation signal $u_{\mathrm{eb}}(k)$ serves as input to the time and frequency envelope shaping blocks which can reconstruct the envelope characteristics of the EB components. Thus the minimum requirement to $u_{\mathrm{eb}}(k)$ is *spectral flatness* and a correct *harmonic structure*.

In the proposed algorithm $u_{\mathrm{eb}}(k)$ is estimated using the fractional pitch lag $\tilde{N}_0 = 3N_0 +$

$N_{0,\mathrm{frac}}$ as well as the adaptive and fixed codebook gains (g_a and g_f) from the G.729A+ decoder. The excitation generation principle is to run the LTP synthesis filter at an increased sampling frequency of 48 kHz (time index k'), thereby producing harmonics of the pitch frequency also in the EB frequency range; the fixed codebook contribution is replaced by scaled noise $n(k')$:

$$\hat{u}_\mathrm{eb}(k') = g_\mathrm{a}\hat{u}_\mathrm{eb}(k' - 2\tilde{N}_0) + g_\mathrm{f}n(k'). \quad (1)$$

The output $\hat{u}_\mathrm{eb}(k')$ of the LTP synthesis filter is bandpass filtered and decimated to produce $u_\mathrm{eb}(k)$ at the desired 16 kHz sampling rate.

3.2 Time Envelope Shaping

The time envelope of the EB signal components is shaped by a scalar factor g_T that is applied to each sample of the excitation signal:

$$u'_\mathrm{eb}(k) = g_\mathrm{T}(k)\, u_\mathrm{eb}(k). \quad (2)$$

To determine $g_\mathrm{T}(k)$, the excitation signal $u_\mathrm{eb}(k)$ is segmented and analyzed in the same manner as done for the original EB signal in the transmitter (Sec. 2). The ratio between the decoded target power $\hat{T}_n$ and the analysis result T_n^R leads to the desired gain factor for the n-th signal segment:

$$g'_\mathrm{T}(n) = \sqrt{\hat{T}_n / T_n^\mathrm{R}}. \quad (3)$$

The final gain function is determined by placing unit pulses, weighted by the respective gain factors from (3), into the middle of each subframe. This sequence of pulses is then interpolated using Hann windows to obtain $g_\mathrm{T}(k)$.

3.3 Frequency Envelope Shaping

Here the input signal to the frequency envelope shaping block, $u'_\mathrm{eb}(k)$, is analyzed and the frequency envelope information $F_{m,\nu}^\mathrm{R}$ is extracted (see Sec. 2). This envelope is compared to the decoded envelope information $\hat{F}_{m,\nu}$. Thus, for each subband (index ν) of the frequency envelope representation a correction gain is determined:

$$G_\mathrm{F}(m,\nu) = \sqrt{\hat{F}_{m,\nu} / F_{m,\nu}^\mathrm{R}}. \quad (4)$$

With the subband gains from (4) a set of FIR filter coefficients $h_\mathrm{F}(k)$ is computed for each signal segment. So the output of the frequency envelope shaping block is:

$$\tilde{s}_\mathrm{eb}(k) = \sum_\kappa u'_\mathrm{eb}(k-\kappa)\, h_\mathrm{F}(\kappa). \quad (5)$$

Finally, the wideband speech signal $\tilde{s}_\mathrm{wb}(k)$ is given by adding $\tilde{s}_\mathrm{eb}(k)$ with $\hat{s}_\mathrm{nb}(k)$.
The coefficients $h_\mathrm{F}(k)$ are determined for each 10 ms signal segment using a weighted sum of prototype bandpass filters $h_\mathrm{F}^{(\nu)}(k)$, i.e.,

$$h_\mathrm{F}(k) = \sum_{\nu=0}^{N_\mathrm{sb}-1} G_\mathrm{F}(m,\nu)\, h_\mathrm{F}^{(\nu)}(k). \quad (6)$$

The subband prototype filters are given by the modulation of a single lowpass prototype $h_\mathrm{lp}(k)$ which is derived from a normalized Kaiser window of length 65 with $\beta = 5$. All of the filters have linear phase with a delay of 32 samples (2 ms). The individual subband frequency responses of the resulting filterbank are shown in Fig. 2. For the case $G_\mathrm{F}(m,\nu) \equiv 1$ the dashed line shows the response of $h_\mathrm{F}(k)$.

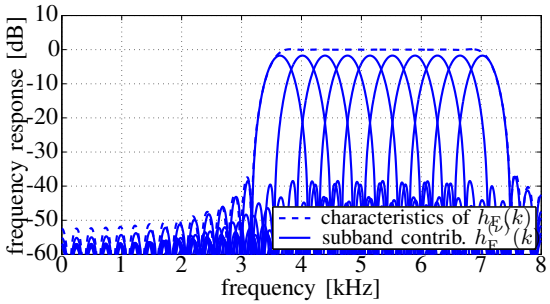

Fig. 2: Frequency responses of $h_\mathrm{F}(k)$ and its subband contributions $h_F^{(\nu)}(k)$.

4. Discussion

In contrast to existing BWE methods, our scheme does not transmit ready-to-be-used gain factors but rather a parametric signal description. Applicable gain factors are computed *at the receiver* and the excitation signal can be analyzed and corrected accordingly. Hence, our scheme is robust against distortions of the excitation signal. *Separate transmission* of time and frequency envelope yields a good resolution in both time and frequency domain and thus a concise reproduction of stationary as

well as transient signals. The *frequency domain shaping* is performed by *FIR filters* instead of all-pole LPC synthesis filters. This avoids typical artifacts like *filter ringing*, which stem from the adaptive switching of all-pole filter characteristics. Our scheme also provides a very *modular concept*, as single blocks in the receiver can easily be exchanged and tuned without any need to alter the transmitter or the bitstream format.

5. Performance Evaluation

To investigate the performance of our new approach, we implemented the BWE algorithm in combination with the G.729A+ speech codec. We conducted listening tests to compare the subjective quality of our approach with that of well-known speech codecs. Reference codecs were the AMR-WB at 8.85 and 12.65 kbit/s, the ITU-T G.722 at 48 kbit/s, and the G.729A codec at 8 kbit/s. Our algorithm has been parameterized for an "add-on bit rate" of 2 kbit/s, i.e., a total bit rate of 14 kbit/s.
We performed a MUSHRA-style test according to ITU-R BS.1534-1 [9] with 15 experienced listeners. The test was conducted with English and German clean speech signals. English speech with additional noise at an SNR of 15 dB was also among the test files. The test results are illustrated in Fig. 3.

6. Conclusion

We have proposed a new embedded wideband coding concept based on BWE with side information on top of a standard narrowband speech codec. Despite its conceptual simplicity the new algorithm results in a natural sounding speech quality. Although presented here for an enhanced G.729A codec, our scheme can be applied to a wide range of existing speech codecs, e.g., [8].
Together with the bitrate scalability of the PCA based quantization scheme, it gives a very high flexibility. Therefore, and since the BWE "add-on" on top of the narrowband codec yields significant quality improvements for comparatively low additional bitrates, our concept is

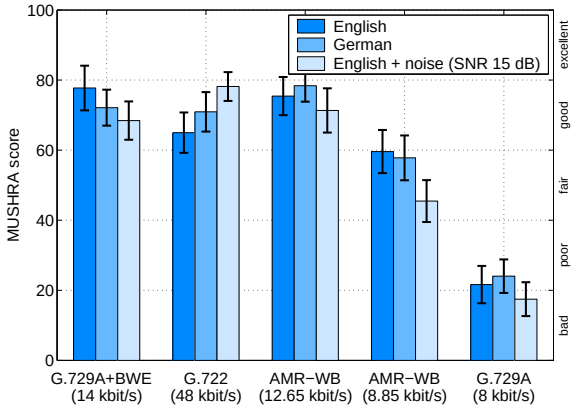

Fig. 3: Results of the subjective listening test with 95% confidence intervals.

a candidate for a large number of interesting applications and was successfully included in the Siemens-Matsushita-Mindspeed candidate submitted to the ITU-T G.729EV competition.

References

[1] ITU-T Recommendation G.729 Annex A, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP). Annex A: Reduced complexity 8 kbit/s CS-ACELP speech codec," 1996.

[2] 3GPP TS 26.190, "AMR wideband speech codec; transcoding functions," Dec. 2001.

[3] P. Jax, "Bandwidth extension for speech," in *Audio Bandwidth Extension*, E. Larsen and R. M. Aarts, Eds., chapter 6, pp. 171–236. Wiley and Sons, Nov. 2004.

[4] H. Carl and U. Heute, "Bandwidth enhancement of narrow-band speech signals," in *Proc. of EUSIPCO*, Edinburgh, Scotland, Sept. 1994, vol. 2, pp. 1178–1181.

[5] 3GPP TS 26.290, "Extended AMR wideband codec; transcoding functions," Sept. 2004.

[6] 3GPP TS 26.404, "Enhanced aacPlus general audio codec; encoder specification; spectral band replication (SBR) part," Sept. 2004.

[7] A. McCree, "A 14 kb/s wideband speech coder with a parametric highband model," in *Proc. of ICASSP*, Istanbul, Turkey, June 2000, vol. 2, pp. 1153–1156.

[8] P. Jax, B. Geiser, S. Schandl, H. Taddei, and P. Vary, "An embedded scalable wideband codec based on the GSM EFR codec," in *Proc. of ICASSP*, Toulouse, France, May 2006.

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