Artificial Bandwidth Extension without Side Information for ITU-T G.729.1

Bernd Geiser [1], *Hervé Taddei* [2], *and Peter Vary* [1]

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

[2]Siemens Networks GmbH
Otto-Hahn-Ring 6
81739 Munich, Germany
herve.taddei@siemens.com

10.21437/Interspeech.2007-663

Abstract

This paper discusses a potential extension of the ITU-T G.729.1 speech and audio codec. The G.729.1 coder is hierarchically organized, i.e., the obtained quality increases with the amount of bits that is received for each frame. In particular, the bit rates of 8 and 12 kbit/s offer *narrowband* (50 Hz – 4 kHz) speech transmission. With a received bit rate of at least 14 kbit/s, the output bandwidth is increased to the *wideband* frequency range (50 Hz – 7 kHz). Here, we investigate efficient methods to provide the full wideband frequency bandwidth already for the lower bit rates of 8 and 12 kbit/s while maintaining interoperability with the standard implementation of G.729.1. These techniques are not necessarily limited to G.729.1 and thus may serve in other applications as well.

Index Terms: speech coding, bandwidth extension

1. Introduction

With the intention of providing a versatile and interoperable codec for wideband speech telephony and audio transmission in Voice over IP (VoIP) networks, ITU-T SG16 standardized Recommendation G.729.1 [1, 2] in May 2006.

The G.729.1 speech and audio coder provides *embedded* coding with 12 bit rates between 8 and 32 kbit/s. The key feature of G.729.1 is its hierarchical bitstream organization: The more of the bitstream "layers" (in total 12) are available to the decoder, the better is the obtained speech or audio quality. In contrast to other multi-mode speech codecs such as the Adaptive Multi Rate (AMR) or Adaptive Multi Rate Wide Band (AMR-WB) codecs, a *bit rate adaptation* is performed by simple *bitstream "truncation"*, i.e., by dropping one or more layers from the hierarchical bitstream. This operation can be performed anywhere in the network according to the capabilities of the transmission link and according to current traffic conditions. Note that a dedicated feedback channel to the encoder is superfluous. Thus, G.729.1 significantly simplifies access to heterogeneous communication networks.

G.729 (with its annexes A and B) is one of the most widely deployed codecs in today's VoIP equipment. In order to facilitate the introduction of wideband conversational services, the main requirement for G.729.1 was its interoperability with G.729 (and its annexes A and B). For this reason, the 8 kbit/s *core layer* of G.729.1 is fully compatible with G.729 ensuring a basic conversational quality. The G.729.1 design philosophy features another narrowband codec mode at 12 kbit/s. It offers an improved narrowband speech quality, especially for interconnections with traditional fixed communication networks such as ISDN (Integrated Services Digital Network) and POTS (Plain Old Telephone System). With the availability of the third bitstream layer (total bit rate: 14 kbit/s), a wideband signal (50 Hz – 7 kHz) can be synthesized using a *bandwidth extension* (BWE) technique [3]. The application to the G.729.1 codec required a few modifications and the respective technology has been labeled "Time Domain Bandwidth Extension" (TDBWE). For codec modes above 14 kbit/s, the wideband signal is refined using an MDCT transform domain algorithm (TDAC, cf. [1]).

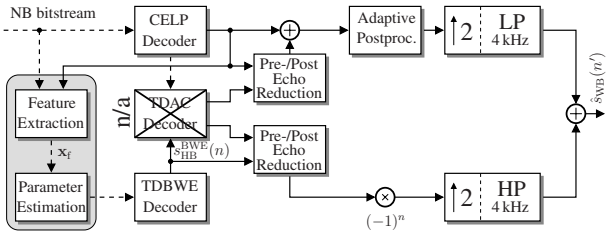

Figure 1: Modified G.729.1 decoder with integrated artificial bandwidth extension (8 and 12 kbit/s modes). The shaded blocks are newly introduced. (TDAC: Time Domain Aliasing Cancellation, TDBWE: Time Domain Bandwidth Extension)

In this paper, we propose an extension to G.729.1 that enables wideband speech synthesis already at bit rates of 8 and 12 kbit/s. Usually, one important cause for objectionable artifacts in the decoded signal is the hard switch of the decoded bandwidth when the bit rate is varied between 12 and 14 kbit/s. The proposed extension mitigates such effects as a constant decoded bandwidth is provided over the whole bit rate range of G.729.1 while full interoperability with the standard is preserved. This interoperability can in principle be achieved by virtually all *estimation based* methods for BWE of speech signals (i.e., without side information), e.g., [4]. We will refer to such methods as *artificial* bandwidth extension (ABWE) in the following. The very specific setup that we consider allows to exploit various synergy effects. In particular, G.729.1 already implements a high quality BWE algorithm (TDBWE) which shall be reused for the problem at hand. Furthermore, we do not only have the synthesized narrowband speech signal available at the decoder but also the respective *parameters* from the 8 or 12 kbit/s bitstream. In fact, these parameters may be exploited for ABWE. Recognizing the described synergies, a very efficient solution for ABWE on top of the narrowband codec layers is feasible (see also [5]). Our modified G.729.1 decoder is illustrated in Fig. 1. The basic idea is to implement an *estimator* for the TDBWE parameters which are not available at 8 and 12 kbit/s modes. Therefore, we (partially) reuse features that are available in the narrowband (8 and/or 12 kbit/s) bitstream.

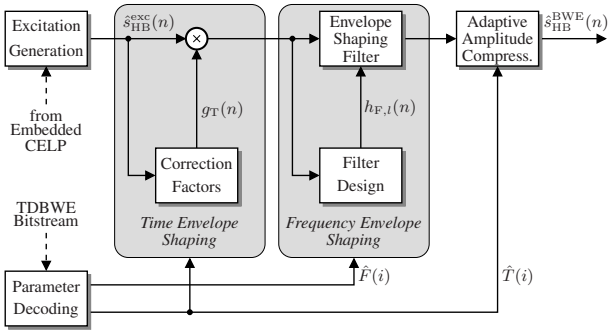

Figure 2: Signal flow of the G.729.1 decoder side bandwidth extension algorithm (TDBWE). Note that the "TDBWE Bitstream" is replaced with an estimated parameter set in our modified system (see Fig 1).

First, we briefly review the TDBWE algorithm (Sec. 2). Then, in Sec. 3, an appropriate set of features is introduced and its eligibility for the estimation of the TDBWE parameters is investigated. The estimation procedure itself is described in Sec. 4 and the obtained results are summarized in Sec. 5. We finally discuss some potential applications as well as possible improvements.

2. TDBWE Revisited

In the TDBWE part of the G.729.1 encoder (14 kbit/s mode), a fairly coarse parametric description of the high frequency components (4 – 7 kHz) of the 20 ms input signal frames is computed. The respective parameter set comprises temporal and spectral energy envelopes. It is quantized with a bit rate of 1.65 kbit/s. At the decoder side, first, a so called "excitation signal" is synthetically generated based on information from the narrowband layers of G.729.1. Then, its time and frequency envelopes are consecutively shaped by gain manipulations and filtering operations to match the transmitted parametric description. Finally, a post-processing procedure attenuates residual artifacts. The TDBWE decoder is shown in Fig. 2. A complete description of this algorithm is provided in the ITU-T G.729.1 recommendation [1].

For the desired application of the TDBWE decoder in an ABWE scheme, the time and frequency envelopes need to be *estimated*[1]. For each 20 ms input speech, the *reference* parameter vectors are given by:

$$\begin{aligned} \mathbf{T} &\doteq (T(0),\ldots,T(15))^{\mathrm{T}} \\ &\text{and} \\ \mathbf{F} &\doteq (F(0),\ldots,F(11))^{\mathrm{T}}. \end{aligned}$$

Thereby, $T(0),\ldots,T(15)$ constitute the logarithmic energies of 16 subframes (1.25 ms each), while $F(0),\ldots,F(11)$ specify the spectral envelope in terms of 12 logarithmic subband energies. In G.729.1, quantized versions of these parameters suffice to synthesize a high-quality subband speech signal.

[1]Note that the *excitation signal generator* in Fig. 2 solely relies on information which is available from the 8 or, optionally, 12 kbit/s bitstream layers of G.729.1. Hence, no additional parameter estimation is required for the excitation generator.

3. Narrowband Features

For the estimation of the TDBWE parameter vectors $\mathbf{F}$ and $\mathbf{T}$, a relevant *feature vector* $\mathbf{x}_{\mathrm{f}}$ has to be chosen which shares sufficient mutual information with $\mathbf{F}$ and $\mathbf{T}$. Investigations about a proper choice for $\mathbf{x}_{\mathrm{f}}$ in ABWE systems have been published in [6]: In terms of mutual information with the high band spectral envelope, the *autocorrelation coefficients* of the narrowband speech performed best, while in terms of class separability, an advantage has been found for the *mel-frequency cepstral coefficients*. Yet, such features are not immediately at hand in the G.729.1 coder and their computation would consume additional complexity. Thus, we choose the also well performing *line spectral pairs* (LSPs) from the CELP core layer.

Further, it has to be taken into account that the TDBWE algorithm does not only comprise a *frequency envelope* but also a *time envelope*. Due to a lack of temporal resolution, the LSPs can not be considered sufficient for the respective estimation. More detailed temporal information is contained in the *gains* of the CELP codebooks of the 8 and 12 kbit/s layers; though, it is still not as fine-grained as the TDBWE time envelope. Hence, we additionally extract a time envelope from the low band (50 Hz – 4 kHz) signal $s_{\mathrm{nb}}(n)$ and include this in the feature vector. This computation only requires little additional complexity. Summarizing, the feature vector $\mathbf{x}_{\mathrm{f}}$ in this study is computed every 20 ms (frame size of G.729.1) and is given by

$$\mathbf{x}_{\mathrm{f}} \doteq (\hat{\mathbf{q}}^{\mathrm{T}}, \mathbf{T}_{\mathrm{nb}}^{\mathrm{T}})^{\mathrm{T}}, \tag{1}$$

where $\hat{\mathbf{q}}$ is the quantized LSP vector of the core layer CELP codec and $\mathbf{T}_{\mathrm{nb}} \doteq (T_{\mathrm{nb}}(0),\ldots,T_{\mathrm{nb}}(15))^{\mathrm{T}}$ is the low band time envelope which is computed in the same manner as the high band time envelope in the TDBWE encoder. In particular, $\mathbf{T}_{\mathrm{nb}}$ comprises the energies of sixteen 1.25 ms subframes.

3.1. Eligibility of the Feature Vector

Here, we assess the applicability of the feature vector $\mathbf{x}_{\mathrm{f}}$ for the estimation of the TDBWE parameters $\mathbf{T}$ and $\mathbf{F}$. For brevity, the following description only focuses on the frequency envelope parameter set $\mathbf{F}$. Nevertheless, it is also valid for the time envelope $\mathbf{T}$.

It is well known that sufficient *mutual information* between the features and the estimation quantity is a necessary condition for concise results, cf. [6] and [7]. Thus, we measured the mutual information $I(\mathbf{x}_{\mathrm{f}};\mathbf{F})$ that is shared between the respective variables. Additionally, like in [7], the high band (4 – 7 kHz) *certainty* that is conveyed by $\mathbf{x}_{\mathrm{f}}$ is computed as the ratio $I(\mathbf{x}_{\mathrm{f}};\mathbf{F})/H(\hat{\mathbf{F}})$. Thereby, $H(\hat{\mathbf{F}})$ constitutes the *entropy* of a uniformly *quantized* parameter vector. This quantization has to be introduced because the discrete Shannon entropy is not applicable to continuous quantities. In fact, we approximate $H(\hat{\mathbf{F}})$ by the *differential entropy* $h(\mathbf{F})$, see [7, 8]:

$$H(\hat{\mathbf{F}}) \approx h(\mathbf{F}) - \log_2(\Delta^{\dim(\mathbf{F})}), \tag{2}$$

where Δ is the quantizer step size. Here, we set Δ such that the obtained mean square error (MSE) $D = \dim(\mathbf{F}) \cdot \Delta^2/12$ is equal to the MSE of the G.729.1 TDBWE quantizer.

The actual measurements of $I(\mathbf{x}_{\mathrm{f}};\mathbf{F})$ and $h(\mathbf{F})$ have been carried out using "k-nearest neighbor statistics" [9, 10] where we set $k = 1$ to achieve minimum bias. These methods are well suited to high-dimensional vector spaces while being data-efficient, i.e., only a comparatively small training set is required.

Table 1: Measurements of mutual information $I(\mathbf{x}_\mathrm{f};\cdot)$, entropies $H(\cdot)$, and "high band certainty" $I(\mathbf{x}_\mathrm{f};\cdot)/H(\cdot)$. The given tolerances specify 95% confidence intervals.

$I(\mathbf{x}_\mathrm{f};\mathbf{F})$	$2.8111 \pm 0.0967\,\mathrm{bit/frame}$
$h(\mathbf{F})$	$23.1795 \pm 0.2442\,\mathrm{bit/frame}$
$H(\hat{\mathbf{F}})$	$\approx 11.1670\,\mathrm{bit/frame}$
$I(\mathbf{x}_\mathrm{f};\mathbf{F})/H(\hat{\mathbf{F}})$	≈ 0.2517
$I(\mathbf{x}_\mathrm{f};\mathbf{T})$	$2.2836 \pm 0.1067\,\mathrm{bit/frame}$
$h(\mathbf{T})$	$30.1008 \pm 0.5810\,\mathrm{bit/frame}$
$H(\hat{\mathbf{T}})$	$\approx 15.5619\,\mathrm{bit/frame}$
$I(\mathbf{x}_\mathrm{f};\mathbf{T})/H(\hat{\mathbf{T}})$	≈ 0.1467

Unfortunately, our choice of k inevitably increases statistical errors in the estimation of $I(\mathbf{x}_\mathrm{f};\mathbf{F})$. Thus, we repeated the experiments several times with different speech samples and computed *confidence intervals* for the obtained results. In particular, we chose $1.2 \cdot 10^5$ active speech frames (20 ms length) from the NTT corpus which were divided into 12 speech samples with 10^4 frames each. For the actual experiment, we used the floating point implementation of G.729.1 (standardized as Annex B). The respective results, both for time and frequency envelopes, are tabulated in Tab. 1.

The measurements of $I(\mathbf{x}_\mathrm{f};\mathbf{F})$ and of the "certainty" values $I(\mathbf{x}_\mathrm{f};\mathbf{F})/H(\hat{\mathbf{F}})$ from Tab. 1 give considerably larger numerical values than previously reported measurements [6, 7, 11]. We can state three reasons for this:

- With a 4 kHz cutoff frequency, the low frequency band is *wider* than in previous investigations, i.e., the obtained features are much more significant for ABWE.
- Conversely, the high band is *narrowed* from the range of $3.4 - 7$ kHz to $4 - 7$ kHz. This, of course, eases the estimation task.
- Compared to previous work, we required a somewhat relaxed distortion constraint D for the entropy estimates. According to (2), this decreases the estimated values of $H(\hat{\mathbf{F}})$. Though, our choice of D (equal to the distortion of the standardized TDBWE quantizer) is justified as the 14 kbit/s mode of G.729.1 has been shown to provide a high speech quality.

Surprisingly, the certainty of the TDBWE time envelope $I(\mathbf{x}_\mathrm{f};\mathbf{T})/H(\hat{\mathbf{T}})$ is comparatively low. An explanation can be its rather high temporal resolution of 1.25 ms which leads to increased entropy $H(\hat{\mathbf{T}})$. Since a smoothed time envelope is often sufficient for ABWE, we repeated the experiment for an *averaged* time envelope per frame, i.e., $\bar{T} = 1/16 \cdot \sum_{i=0}^{15} T(i)$. The respective measurements yield:

$$I(\mathbf{x}_\mathrm{f};\bar{T}) = 1.5719 \pm 0.0414\,\mathrm{bit/frame}$$
and
$$H(\bar{\hat{T}}) = 4.2999 \pm 0.0666\,\mathrm{bit/frame},$$

which gives a certainty for $\bar{T}$ of approximately 0.3656.

Motivated by these results, we expect that the intended ABWE scheme for the G.729.1 TDBWE parameters can perform satisfactory. Next, we will describe our example implementation and then present some performance and quality evaluation results.

4. High Band Parameter Estimation

For this study we designed minimum mean square error (MMSE) estimators for the TDBWE parameter vectors $\mathbf{F}$ and $\mathbf{T}$. Our implementation is based on the framework from [4] which is briefly summarized in the following. Therefore, we use a "generic" estimation quantity $\mathbf{y}$ instead of the actual vectors $\mathbf{F}$ and $\mathbf{T}$. In addition to the MMSE estimation, we implemented a simple *post-processing* of the estimated parameters as described in Sec. 4.3.

4.1. MMSE Estimation

First, as a prerequisite, we define a sequence of feature vectors as $\mathbf{X}_\mathrm{f} \doteq \{\mathbf{x}_\mathrm{f}(1), \ldots, \mathbf{x}_\mathrm{f}(m)\}$, where the numbers 1 to m specify the respective frame indices. In particular, m designates the *current* frame, i.e., $\mathbf{x}_\mathrm{f}(m) = \mathbf{x}_\mathrm{f}$.

The criterion for MMSE estimation of a vector $\mathbf{y}$ with given observations $\mathbf{X}_\mathrm{f}$ is $E\left\{||\mathbf{y} - \tilde{\mathbf{y}}||^2 \middle| \mathbf{X}_\mathrm{f}\right\} = \min$, where $\tilde{\mathbf{y}}$ is the respective estimate. The solution to this optimization problem is the conditional expectation $\tilde{\mathbf{y}}_\mathrm{MMSE} = E\left\{\mathbf{y} \middle| \mathbf{X}_\mathrm{f}\right\}$, cf. [4]. Using a precomputed codebook $\mathcal{C} = \{\hat{\mathbf{y}}_\mathbf{1}, \ldots, \hat{\mathbf{y}}_\mathbf{M}\}$ for the vectors $\mathbf{y}$ (e.g., obtained with the LBG algorithm [12]), this MMSE estimate can be expressed as

$$\tilde{\mathbf{y}}_\mathrm{MMSE} = \sum_{\hat{\mathbf{y}}_i \in \mathcal{C}} \hat{\mathbf{y}}_i \cdot P(\hat{\mathbf{y}}_i | \mathbf{X}_\mathrm{f}) \tag{3}$$

which essentially is a weighted sum over the M centroids of the codebook $\mathcal{C}$. Thereby, the weights $P(\hat{\mathbf{y}}_i|\mathbf{X}_\mathrm{f})$ specify *a posteriori probabilities*.

4.2. A Posteriori Probabilities

The probabilities $P(\hat{\mathbf{y}}_i|\mathbf{X}_\mathrm{f})$ from (3) can be reformulated as

$$P(\hat{\mathbf{y}}_i|\mathbf{X}_\mathrm{f}) = \frac{p(\hat{\mathbf{y}}_i, \mathbf{X}_\mathrm{f})}{p(\mathbf{X}_\mathrm{f})} = \frac{p(\hat{\mathbf{y}}_i, \mathbf{X}_\mathrm{f})}{\sum\limits_{\hat{\mathbf{y}}_i \in \mathcal{C}} p(\hat{\mathbf{y}}_i, \mathbf{X}_\mathrm{f})}. \tag{4}$$

The joint densities $p(\hat{\mathbf{y}}_i, \mathbf{X}_\mathrm{f})$ in (4) are computed as the product of the so-called *observation densities* $p(\mathbf{x}_\mathrm{f}\big|\hat{\mathbf{y}}_i)$ of the *current* feature vector $\mathbf{x}_\mathrm{f}$ and of specific joint densities that comprise accumulated *a priori knowledge* [4]:

$$p(\hat{\mathbf{y}}_i, \mathbf{X}_\mathrm{f}) = p(\mathbf{x}_\mathrm{f}\big|\hat{\mathbf{y}}_i) \cdot p(\hat{\mathbf{y}}_i, \mathbf{X}'_\mathrm{f}), \tag{5}$$

where $\mathbf{X}'_\mathrm{f} \doteq \{\mathbf{x}_\mathrm{f}(1), \ldots, \mathbf{x}_\mathrm{f}(m-1)\} = \mathbf{X}_\mathrm{f} \setminus \mathbf{x}_\mathrm{f}$. For our estimation scheme, we approximate the observation densities $p(\mathbf{x}_\mathrm{f}\big|\hat{\mathbf{y}}_i)$ with Gaussian Mixture Models (GMMs) during an offline training phase. Then, given an actual observation $\mathbf{x}_\mathrm{f}$, these GMMs are used to compute the values of $p(\mathbf{x}_\mathrm{f}\big|\hat{\mathbf{y}}_i)$ for all $i \in \{1, \ldots, M\}$. The computation of the a priori term $p(\hat{\mathbf{y}}_i, \mathbf{X}'_\mathrm{f})$ exploits *a priori knowledge* of first order, i.e., *state transitions* are explicitly considered (see [4]). Finally, insertion of (5) into (4) and then (3) yields the desired MMSE estimate $\tilde{\mathbf{y}}_\mathrm{MMSE}$ for the vector $\mathbf{y}$.

4.3. Post-Processing of the Estimated Parameters

The speech quality can be improved by using a smoother temporal evolution of the estimated parameters. Thus, we apply a short gliding average to the estimated vectors (see also [13]):

$$\tilde{\mathbf{y}} = \frac{1}{2} \cdot \left(\tilde{\mathbf{y}}_\mathrm{MMSE} + \tilde{\mathbf{y}}_\mathrm{MMSE,old}\right), \tag{6}$$

where $\tilde{\mathbf{y}}_\mathrm{MMSE,old}$ is the MMSE estimate that has been obtained for the preceding frame. In effect, this procedure attenuates artifacts that stem from strongly time-variant estimates.

5. Results

We will now exemplarily show the estimation performance that can be obtained with the described framework based on 6 min of clean speech from the NTT corpus. In addition, we judge the resulting wideband speech quality.

For the present tests, we configured the MMSE estimators from Sec. 4 with 7 bit codebooks for $\mathbf{T}$ and $\mathbf{F}$ that have been trained via [12]. The GMMs representing the observation densities $p(\mathbf{x}_\mathrm{f}|\hat{\mathbf{y}}_i)$ comprise 8 mixture components each.

5.1. Estimation Performance

Figure 3 visualizes the performance of the parameter estimation. As an example, we show the true and estimated values of the high band *time envelope* for a short utterance of a female English speaker. For the most part, the estimate follows the true envelope rather accurately, even for sounds of high energy (e.g., fricatives). Though, the estimate lacks some temporal fine structure which is coherent with the results from Sec. 3.1.

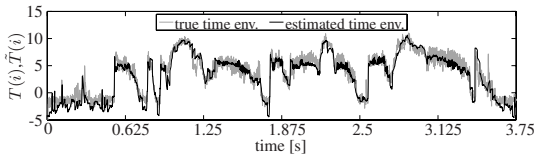

Figure 3: Example for the estimation of the time envelope.

In order to compute the achieved high band *spectral* distortion, we used 8th order AR models of the true and of the estimated high band signals[2]. With this method, a high band *root mean square log spectral distortion* per frame of $\bar{d}_\mathrm{LSD} \approx 5.05\,\mathrm{dB}$ was measured. Moreover, for reference, the corresponding value has also been obtained for the *quantized* TDBWE output signal. The respective measurement yields $\bar{d}_\mathrm{LSD,quant.} \approx 3.55\,\mathrm{dB}$.

5.2. Subjective Speech Quality

Subjectively, we can conclude that the proposed ABWE scheme is capable of producing a high band signal with few artifacts only. Overall, our method can regenerate the impression of a real wideband signal. Naturally, when compared with a *coded* version, we note that the sound is still a little "muffled" and lacks some brilliance. We believe that this is due to the inevitably limited dynamic range in the estimate of the frequency envelope. However, certain phonemes that are sometimes problematic for ABWE algorithms, especially */s/* and */f/*, seem to be identified correctly in most cases. We attribute this to the comparatively high band split frequency of 4 kHz which significantly increases the mutual information measure (see Sec. 3.1).

6. Discussion

We can identify several potential applications for an ABWE scheme within the framework of the G.729.1 codec:

- In general, the speech quality and intelligibility can be increased for the 8 and 12 kbit/s modes of G.729.1.
- In a heterogeneous conference scenario using G.729.1, there might be users who perceive a *mix* of narrow- and wideband speech which impairs the conversational quality. Here, ABWE can provide a constant bandwidth.
- Bit rate switchings due to network congestion may occur between bit rates of, e.g., 12 and 14 kbit/s. Normally, this causes severe artifacts as the acoustical bandwidth is significantly narrowed during a short period. Appropriately inserting the estimated high band signal can virtually remove such artifacts.

We further note that the obtained wideband speech quality for the 8 and 12 kbit/s codec modes can be greatly increased if we allow to transmit some compact side information (e.g., resiual estimation error) using CELP data hiding techniques [14]. Finally, for the estimation part, we did not yet implement methods to reduce the feature dimension (e.g., *linear discriminant analysis*). Such techniques can significantly reduce the computational complexity of the parameter estimation while at least preserving the estimation performance [6].

7. Conclusions

In this paper we have proposed a method for artificial bandwidth extension of speech on top of the 8 and 12 kbit/s modes of the ITU-T G.729.1 codec. Thereby, we reused the available narrowband parameters as features for an estimation task. The actual bandwidth extension is performed via the standardized TDBWE algorithm of the codec. Hence, the required additional computational complexity is entirely determined by the estimation procedure. Moreover, measurements of mutual information and "high band certainty" have shown that the TDBWE parameters are actually amendable to estimation. Using an exemplary implementation we have found that satisfactory wideband speech quality can be obtained. Especially the particularly critical fricative sounds can be identified correctly in most cases. The method can, of course, also be useful for other codecs.

8. References

[1] ITU-T Rec. G.729.1, "G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729."

[2] S. Ragot *et al.*, "ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and Voice over IP," in *Proc. of IEEE ICASSP*, Honolulu, Hawai'i, USA, Apr. 2007.

[3] 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 IEEE ICASSP*, vol. 1, Toulouse, France, May 2006, pp. 5–8.

[4] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.

[5] M. M. Deshpande and K. R. Ramakrishnan, "A novel BWE scheme based on spectral peaks in G.729 compressed domain," in *Proc. of EUSIPCO*, Antalya, Turkey, Sept. 2005.

[6] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals," in *Proc. of IEEE ICASSP*, vol. 1, Montreal, Canada, May 2004, pp. 697–700.

[7] M. Nilsson, H. Gustafsson, S. V. Andersen, and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech," in *Proc. of IEEE ICASSP*, vol. 1, Orlando, FL, USA, May 2002, pp. 525–528.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[9] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69 (066138), 2004.

[10] L. F. Kozachenko and N. N. Leonenko, "A statistical estimate for the entropy of a random vector," *Probl. Information Transmission*, vol. 23, no. 2, 1987.

[11] A. H. Nour-Eldin, T. Z. Shabestary, and P. Kabal, "The effect of memory inclusion on mutual information between speech frequency bands," in *Proc. of IEEE ICASSP*, Toulouse, France, May 2006.

[12] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 28, no. 1, pp. 84–95, Jan. 1980.

[13] U. Kornagel, "Techniques for artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 86, no. 6, pp. 1296–1306, July 2006.

[14] B. Geiser and P. Vary, "Backwards compatible wideband telephony in mobile networks: CELP watermarking and bandwidth extension," in *Proc. of IEEE ICASSP*, Honolulu, Hawai'i, USA, Apr. 2007.

[2] This model order has also been chosen in [4] (although for a wider extension band, i.e., for the frequency range of 3.4 – 7 kHz).