EMBEDDED SPEECH CODING BASED ON PYRAMID CELP

Christoph Erdmann and Peter Vary

Institute of Communication Systems and Data Processing
Aachen University of Technology
Templergraben 55, D-52056 Aachen, Germany
e-mail: erdmann@ind.rwth-aachen.de

ABSTRACT

In this paper we investigate *embedded coding* of speech in the CELP (Code-Excited-Linear-Prediction) framework. Compared to other known approaches of variable bit rate speech coding, such as the Adaptive Multi-Rate (AMR) codec, embedded coding systems allow bit rate reductions at any point along the communication network, without changes of the encoder and the decoder. Thus, the quality of the decoded speech increases with the amount of received bits. Aiming at a coding scheme, that produces such a hierarchically-structured bit stream, we focus on a decomposition of the excitation signal by means of pyramid coding [1]. To achieve reasonable speech compression, Analysis-by-Synthesis (AbS) based quantization of the pyramid layers is designed in a CELP-type fashion, called P-CELP (pyramid CELP)[2]. Besides an efficient design of fixed algebraic codebooks in each pyramid layer, special attention is payed to the integration of an adaptive codebook. To achieve maximum performance, the proposed coding scheme is used for wideband (50 Hz - 7 kHz) speech.

1. INTRODUCTION

Variable bit rate (VBR) coding is particularly advantageous for speech coding applications, where channel impairments play a major role in the design of the coder. The additional degree of freedom by adapting the bit rate to the changing channel, allows it to maintain real-time speech transmission even under network conditions, where fixed rate coders fail to operate. In most VBR coding systems it is common practice that both encoder and decoder work synchronously at the same bit rate. Any changes of the bit rate will be controlled by a rate adaptation unit needing a feedback channel to inform encoder and decoder about the changed bit rate of the transmission system.

However, in some applications, the rate decision is located at a distant point in the network, for example at the conference bridge for telephoneconferencing or the multiplexer for digital circuit multiplication equipments. In this case so-called *embedded coding* is needed, in which a single encoding algorithm generates a fixed-rate hierarchically-structured bit stream from which reduced-rate bit streams can be extracted. This implies a hierarchy of blocks of bits within which are embedded further subblocks. By dropping (sub-)blocks bit rate reductions can be done anywhere along the communication path. Depending on the number of (sub-)blocks received in a finite time interval, the decoder can select the rate and fills in the missing bits with zeros prior to decoding with its fixed decoding algorithm.

Although this hierarchical concept is a natural one for many classes of speech coders, only few examples exist, where the concept of embedded coding was combined with CELP coding [3]. In this paper we propose a new coding scheme based on the design of an algorithm to hierarchically encode the excitation signal of a CELP coder by means of pyramid coding.

2. PYRAMID CODING OF SPEECH SIGNALS

Presently, pyramid coding is much more explored in the field of image coding [1] than for speech coding applications, which led to a far better understanding of pyramid coding in the context of image compression. A so-called *image pyramid* means the representation of an image with multiple resolutions. The different resolution-levels are denoted as pyramid layers. A simple yet important structure of a pyramid is the so-called *Gauss* pyramid $\mathbf{g}$, which is depicted on the left side of Fig. 1.

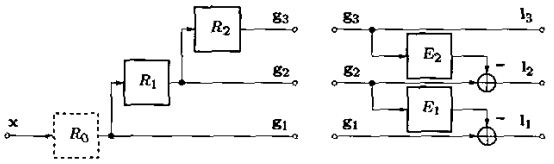

Fig. 1. Three-stage Laplace pyramid $\mathbf{l} = (\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3)$, based on a Gauss pyramid $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3)$.

A complete description of the Gauss pyramid is given by a set of reduce-operators R_i (i.e. lowpass filtering[1] followed by decimation). Applying these reduce-operators sequentially to the input signal $\mathbf{x}$, one obtains the different layers $\mathbf{g}_i$ of the Gauss pyramid.

Using Gauss pyramids is most self-evident, when representing signals that should be applied to scalable compression algorithms. When combined with interpolation the hierarchical structure yields the so-called *Laplace* pyramid.

Based on the decomposition of the signal into a Gauss pyramid, the representation of the Laplace pyramid is derived in a second step by defining a set of expand-operators E_i, which are chosen to be complementary to the reduce-operators R_i (i.e. upsampling followed by interpolation). Applied to the $(i+1)$-th layer $\mathbf{g}_{i+1}$ of the Gauss pyramid, E_i provides a signal mapped to the size of the i-th layer. As shown on the right side of Fig. 1, the expand-operator E_i approximates the Gauss layer $\mathbf{g}_i$ from the less accurate layer $\mathbf{g}_{i+1}$. Thus, each layer $\mathbf{l}_i$ of the Laplace pyramid $\mathbf{l}$ describes the error resulting from the difference between a Gauss layer $\mathbf{g}_i$ and its approximation from the layer $\mathbf{g}_{i+1}$.

A simple rule for reconstructing the original signal $\mathbf{x}$ from its Laplace layers is shown in Fig. 2a) and Fig. 2b), respectively. Since in this paper we rely on the paradigm of hierarchical coding to produce an embedded bit stream, the reconstruction scheme shown in Fig. 2b) is particularly interesting. Each layer $\mathbf{l}_i$ is separatly expanded to a layer $\mathbf{l}_i'$:

$$\mathbf{l}_i' = \overbrace{E_{i-1}}^{\cdots}(\mathbf{l}_i) \quad . \tag{1}$$

[1] Originally, Gauss-type FIR-filters were used for the reduce-operation, which led to the name Gauss pyramid.

0-7803-7549-1/02/$17.00 ©2002 IEEE.

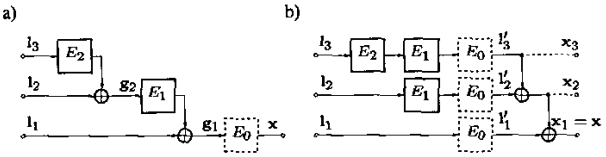

Fig. 2. *a) Reconstruction of three-stage Laplace pyramid*
b) Equivalent reconstruction scheme

The corresponding expand-operation $\ddot{\ddot{E}}_{i-1}$ results from a concatenation of all expand-operators, which are relevant for this layer:

$$\ddot{\ddot{E}}_i = E_0 \circ \cdots \circ E_i, \quad i = 1, \ldots, L-1. \tag{2}$$

Due to this expand procedure all layers $\mathbf{l}'_i$ have the same size as the original signal $\mathbf{x}$. I.e. the $\mathbf{l}'_i$ have the sampling rate of the original signal, but their frequency components are limited to the frequency components of the corresponding Laplace layer $\mathbf{l}_i$. The simple reconstruction rule is given by

$$\mathbf{x} = \sum_{i=1}^{L} \ddot{\ddot{E}}_{i-1}(\mathbf{l}_i) = \sum_{i=1}^{L} \mathbf{l}'_i \quad . \tag{3}$$

The layers of the Laplace pyramid $\mathbf{l}_1, \ldots, \mathbf{l}_L$ feature the required hierarchy to form an embedded bit stream where each layer is coded into a sub-block of bits. With an increasing number of available sub-blocks (i.e. Laplace layers $\mathbf{l}_i$), the decoder can reconstruct a signal $\mathbf{x}_i$, which approximates the original signal $\mathbf{x}$ with increasing accuracy in terms of bandwidth.

In [2], we have already demonstrated, how to achieve scalability in terms of perceptual speech quality, by applying the Laplace pyramid to the residual signal of a predictive speech coder, instead of the speech signal itself. We transmit the linear prediction (LP) synthesis filter coefficients as side information. Thus, the spectral envelope of the wideband speech signal will be properly reconstructed over the whole frequency range, regardless of the accuracy of the reconstructed excitation.

To ensure, that the reconstruction of the Laplace pyramid in the residual domain also yields a wideband signal, regardless of the number of pyramid layers available for decoding, we introduced a modified expand operation, that employs up-sampling but omits the following lowpass-interpolation. This kind of high-frequency regeneration by aliasing [4], places frequency reversed images of the low frequency components into the empty upper frequency range of the up-sampled signal, producing a rough approximation of the true high-frequency spectrum. The additional decoding of the upper layers stepwise corrects this systematic reconstruction error until full accuracy of the original residual is reached by decoding all the pyramid layers.

3. PYRAMID CELP (P-CELP) CODING

In Analysis-by-Synthesis (AbS) based predictive speech coding CELP coders have been successful in obtaining high speech quality mainly due to efficient quantization of the excitation signal. In CELP coding, except for bit rates below 4 kbit/s, the predominant part of the available bit rate is used to encode the excitation. Thus, the exciation is most amenable to be encoded with variable bit rate. In fact most VBR speech coders based on CELP, e.g. GSM-AMR codec, employ quantization of the excitation with different bit rates yielding scalability of the quality of the reconstructed speech along with the employed bit rate.

In a standard CELP coder, the excitation is generated by adding vectors from an adaptive codebook (ACB) and a fixed codebook (FCB). A closed-loop search is sequentially performed in both codebooks with respect to the minimization of the reconstruction error. This search procedure is also referred to as CELP search. The ACB, which contains delayed segments of the past excitation, is assumed to contribute to the periodic component of the excitation, while the FCB is generally expected to handle the innovative (i.e. random) component.

A simple way to achieve CELP coding with embedded property is the sequential usage of multiple fixed codebooks [3]. With each additional codebook vector found, the reconstruction error is successively minimized yielding a hierarchical bit stream. Accordingly, the decoder can sucessiveley refine the reconstructed excitation and thus the reconstructed speech by decoding additional FCB vectors. A major drawback of this approach is, that each codebook vector found aims to refine the reconstruction error for the entire bandwidth. Since the targets of these successive optimizations are not orthogonal, the achieved minimization of the reconstruction error is much worse compared to a joint optimization with one single codebook. Thus, the perceived quality-increment per decoded FCB contribution is poor in proportion to the additionally employed bit rate.

To improve the relation between bit rate and quality-increment we propose to combine CELP search with pyramid coding. In the first step, we formulate an alternative expression to construct the Laplace pyramid $\mathbf{l}'$ by

$$\begin{aligned} \mathbf{l}'_L &= \mathbf{g}'_L = \ddot{\ddot{E}}_{L-1}(\mathbf{g}_L) \\ \mathbf{l}'_i &= \ddot{\ddot{E}}_{i-1}(\mathbf{g}_i) - \sum_{\nu=i+1}^{L} \mathbf{l}'_\nu \quad , \end{aligned} \tag{4}$$

with $\ddot{\ddot{E}}_i$ beeing the concatenated expand-operation from (1). Using this construction rule, separate quantization Q_i of each Laplace layer can be established as shown in Fig. 3. This repre-

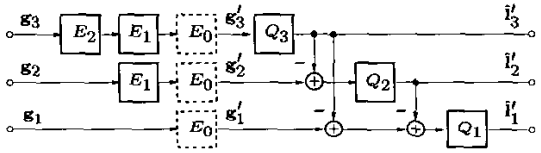

Fig. 3. Three-stage Laplace pyramid with quantization

sentation of a Laplace pyramid with quantization already includes a closed-loop property. Each unquantized Laplace layer $\mathbf{l}'_i$ includes the quantization errors of its preceding layers. Hence, quantization of one layer takes into account the quantization errors already made and possibly corrects them. So far, we have not made any assumptions how the quantization Q_i of each Laplace layer is performed. Since we are aiming at CELP coding, Q_i is realized by closed-loop CELP search. Fig. 4 illustrates the complete encoder structure of the resulting pyramid CELP coder.

3.1. Fixed codebook search

The structure of the fixed codebooks FCB1,..FCB3 corresponds to a state-of-the-art algebraic codebook employing a non-exhaustive tree-search algorithm as proposed in [5]. As required, the target signal $\mathbf{x}'_i$ as well as the reconstructed signal in each layer covers the full (wideband) signal bandwidth. On the other hand, each pyramid layer $\mathbf{g}'_i$ matches the true spectral shape of the excitation $\mathbf{d}$ only in its particular frequency bandwidth it was reduced to. Wideband coverage is achieved upon usage of the expand operation E_i. Through up-sampling the remaining

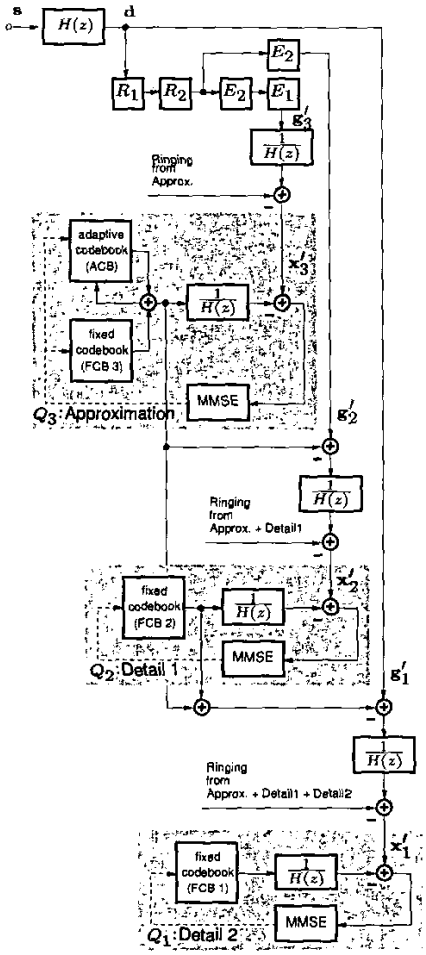

Fig. 4. P-CELP: embedded encoder featuring closed-loop codebook search within a three-stage Laplace pyramid.

frequency range is filled up with regularly repeated images of the reduced frequency spectrum. To exploit the full advantage of this approach, the FCB vectors in each layer have to feature the same spectral repetition-property as their respective pyramid layer $\mathbf{g}_i'$. This can be achieved by the design of special FCB's with restricted pulse positions according to the sampling rate reduction of the pyramid layer they are applied to.

Hence, the minimization of the reconstruction error is limited to the respective frequeny-band which is truly represented by that particular layer. When neglecting the quantization error caused by preceeding layers, the target signals $\mathbf{x}_i'$ for the FCB search in each pyramid layer are orthogonal. Thus, the successive optimizations performed with this coding scheme are nearly independet from each other yielding a more effective minimization of the overall (wideband) reconstruction error. Compared to the performance of conventional embedded CELP proposed in [3], this results in an improved relation between bit rate and quality-increment.

3.2. Adaptive codebook search

Fig. 4 shows a setup, where ACB search is performed only in the base-layer to contribute to the so-called *Approximation*. This is to prevent severe mismatches between the states of encoder and decoder for the case that the upper layers (i.e. Detail1 and Detail2) will not reach the receiver for a longer time. If solely the quantized base-layer, though it produces the coarsest approximation of the excitation, is used to update the ACB, the embedded property can be kept even when the reconstruction only rests upon the Approximation. In fact, this is a fail-save setup. Depending on the average bit rate and the packet loss charateristics of the channel, this setup can be altered to have more pyramid layers feeding the ACB. By extending the ACB search and ACB update to a greater number of pyramid layers, the resulting coder achieves a higher potential speech quality at the expense of a decreased robustness.

3.3. Performance

In [2] we already demonstrated, that the proposed P-CELP coder significantly improves error robustness when transmitting over lossy packet networks, compared to a standard CELP coder.

To demonstrate the improved relation between bit rate increment and quality-increment, we compare the P-CELP performance against a conventional embedded CELP coder as proposed in [3]. Therefore, both coders were designed as wideband coders, including one base-layer (Approximation) and three enhancement-layers (Detail1,..., Detail3) using identical bit rates. The employed bit rates are: Approximation: 6.75 kbit/s + Detail1: 13.95 kbit/s + Detail2: 25.15 kbit/s + Detail3: 38.35 kbit/s.

Informal listening tests show, that for both coders the performance of the approximation is rather poor. Due to the extreme degree of pyramid decomposition at this level the P-CELP coder sounds more synthetic than the conventional embedded CELP coder. At 13.95 kbit/s (+Detail1) the performance of both coders is rated equal. For the higher bit rates the P-CELP clearly outperforms the conventional embedded CELP, the quality of which stagnates at the performance reached at 13.95 kbit/s.

4. CONCLUSIONS

In this paper we presented a new embedded speech coding concept based on pyramid CELP (P-CELP). The coder generates a hierarchically-structured bit stream by using pyramid decomposition of the excitation signal. Each pyramid layer is efficiently quantized by CELP search. The algorithm offers progressive decoding of the speech signal with increasing perceptual speech quality. The proposed algorithm allows bit rate reductions at any point along the communication network, without the encoder and the decoder knowing about theses changes. This behaviour is particularly advantageous when applied to real-time voice transmission over lossy packet networks or broadcast. Compared to conventional embedded CELP speech coders, the proposed P-CELP coder significantly improves the relation between additionally employed bit rate and the achieved quality-increment.

5. REFERENCES

[1] P.J. Burt and E.H. Adelson, "The Laplacian Pyramid as a Compact Image Code", *IEEE Trans. on Comm.*, vol. 31, no. 4, Apr. 1983.

[2] C. Erdmann, D. Bauer, and P. Vary, "Pyramid CELP: Embedded Speech Coding for Packet Communications", in *Proc. ICASSP*, Orlando, Florida, May 2002.

[3] A. Le Guyader, C. Lamblin, and E. Boursicaut, "Embedded Algebraic CELP/VSELP Coders for Wideband Speech Coding", *Speech Communication*, vol. 16, pp. 319–328, 1995.

[4] J. Makhoul and M. Berouti, "High-Frequency Regeneration in Speech Coding Systems,", in *Proc. ICASSP*, Washington, DC, 1979.

[5] C. Laflamme, J.-P. Adoul, H.Y. Su, and S. Morisette, "On Reducing Computational Complexity of Codebook Search in CELP Coder Through the Use of Algebraic Codes,", in *Proc. ICASSP*, Albuquerque, New Mexico, 1990.