A 1.6 KBPS SPEECH CODEC USING SPECTRAL VECTOR QUANTIZATION OF DIFFERENTIAL FEATURES

Hyoung-Gook Kim [1,2], Klaus Obermayer[2],
Mathias Bode[1], Dietmar Ruwisch [1]
[1]Cortologic AG, Berlin, Germany
[2]Department of Computer Science,
Technical University of Berlin, Germany
{kim, bode, ruwisch}@cortologic.com, oby@cs.tu-berlin.de

ABSTRACT: In this paper we propose an efficient algorithm for a low rate and low complexity speech compression algorithm based on spectral vector quantization of differential features in the frequency domain. Speech signals can be effectively encoded at medium transmission rates maintaining high quality of the reconstructed speech. To operate at lower transmission rates with minimized quality drawbacks, we use differential feature vector coding in the frequency domain. From the spectrogram comparison, it is shown that the use of the proposed method provides reasonable quality of synthesized speech.

1. INTRODUCTION

In recent years, very low rate speech coding is becoming an important research area in modern speech coders. Speech signals can be effectively encoded at various transmission rates maintaining high quality of the reconstructed speech. The compression of speech signals has many practical applications. One example is in digital cellular telephony where many users share the same frequency band. Another example is digital voice storage (e.g. answering machines). With current compression techniques, it is possible to reduce the rate to 8 kbps with almost no perceptible loss in quality. Further compression is possible at a cost of lower quality. Almost all of the current low-rate speech coders are based on the principle of Linear Predictive Coding (LPC). However, there are several drawbacks in these speech codecs. A speech wave form is partitioned into many small segments, speech frames, which are classified as "voiced" or "unvoiced". Due to noise interference and ambiguity of the transient state between the voiced and unvoiced, this method cannot achieve a high performance of the classification such that the quality of synthetic speech is degraded. Waveform coders such as Code Excited Linear Prediction (CELP) (Schroeder & Atal, 1985) algorithms have been dominant in speech coding at rates above 4 kbps. However, for rates around 4kbps and below, speech coding in the spectral domain has recently shown potential for better quality than the widely used CELP based codecs. Spectral coding of speech is usually based on harmonic coding (sinusoidal coding) employed in coders like Sinusoidal Transform Coding (STC) (McAulay & Quatieri, 1995), Multiband Excitation Coding (MBE) (Griffin & Lim, 1998) (Cho & Kim, 1997) and Spectral Excitation Coding (Lupini & Cuperman, 1995). Harmonic coders are attractive methods to obtain high quality reconstructed speech by retaining only the spectral harmonic magnitudes and using a synthetic harmonic phase at low bit rates.

In this paper, we propose a 1.6 kbps speech codec for low bit-rate coding. The reduction in rate to 1.6 kbps is accomplished by using a spectral vector quantization method for differential features.

2. ALGORITHM DESCRIPTION

In our implementation (see Fig. 1), the input speech is presented to the encoder in frames of 256 samples with a 50 % overlap. As the first signal processing step, the pitch period is estimated by searching the maximum of the normalized correlation. In the first stage, the pitch is roughly estimated to ensure a smoothed pitch evolution. For each frame the correlation value, the pitch period and the number of periods per frame are determined. In the second stage, the pitch period is adapted by quadratic interpolation. The temporal autocorrelation method is widely used for pitch determination in the time domain. Given a speech signal $s(n)$, the temporal autocorrelation at a candidate pitch τ is defined as

$$R^T(\tau) = \frac{\sum_{n=0}^{N-\tau-1} s(n)s(n+\tau)}{\sqrt{\sum_{n=0}^{N-\tau-1} s^2(n) \sum_{n=0}^{N-\tau-1} s^2(n+\tau)}} \qquad (1)$$

where $s(n)$ is the zero-mean speech signal, and N is the number of samples for pitch determination. These samples along with N samples of the previous frame are multiplied by a smoothed trapezoidal window. These samples are transformed into the frequency domain by a $2N$-point FFT. The length of the FFT is dependent not only on the pitch period but also on voicing information. The FFT spectrum is interpolated for keeping the spectral envelope and the pitch period is reconstructed by quadratic interpolation. The temporal autocorrelation method produces exact pitch values in most cases, however, pitch errors may occur. To compensate for the pitch error, the number of periods per frame is determined by three different methods: the number of periods per frame from the temporal correlation, from the FFT and the old number of periods per frame of the last frame. Given a pitch candidate τ , the optimal integer pitch T is determined as

$$T = \tau \ / C_{PitchPeriod} \qquad (2)$$

where $C_{PitchPeriod}$ is the number of periods per frame. In the frequency domain, the spectral values are grouped together to form 30 band filters referred to as channels and are subjected to an empirical memoryless non-linear function (e.g. power laws). In the frequency domain, an encoding operation is next performed by "subtraction of the last synthesis vector" (see Fig. 2). The synthesis vectors updated by means of the differential codebook vectors are subtracted from feature vectors obtained by the above mentioned non-linear function to yield differential feature vectors. The speech compression algorithm classifies the differential feature vectors by a vector quantizer i.e. by searching the codebook index corresponding to the excitation prototype that minimizes the error. The idea of the "subtraction of synthesis vector" is that the codebook index of each differential feature vector can be transmitted and, from this, the entire synthesis feature vectors can be reconstructed by the "summation of synthesis vectors". As shown in Fig. 1a, the compressed speech information includes the codebook index, the number of periods per frame and the pitch period.

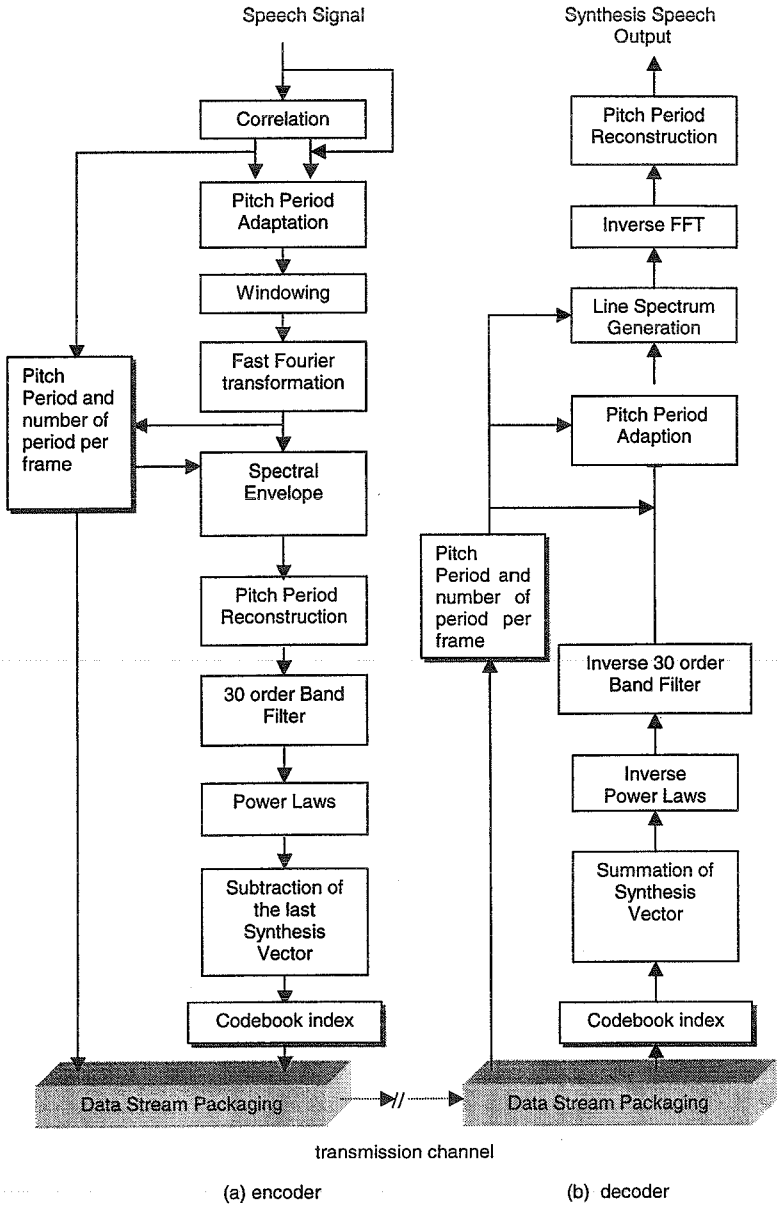

Fig. 1: The encoder and the decoder of the compression algorithm

The decoder uses the same codebook vectors to produce the feature vectors (see Fig. 2). The "summation of synthesis vectors" in the decoder converts the received codebook index into the differential codebook vectors by the inverse vector quantization and produces the synthesis vectors by summing up the differential codebook vectors. The resulting feature vectors are transformed by the inverse power law function. The interpolated spectrum envelope representation is discretized by means of the transmitted pitch period and the number of periods per frame. The synthetic speech output signal is generated by inverse Fourier transform with constant phases and by pitch period reconstruction. This leads to good results in case of voiced as well as unvoiced signals, without distinguishing them explicitly. In case of vowels the pitch period is nearly constant, thus a quasi periodic signal is synthesized. Unvoiced consonants produce stronger fluctuations of the pitch period estimate leading to noise components in the output signal.

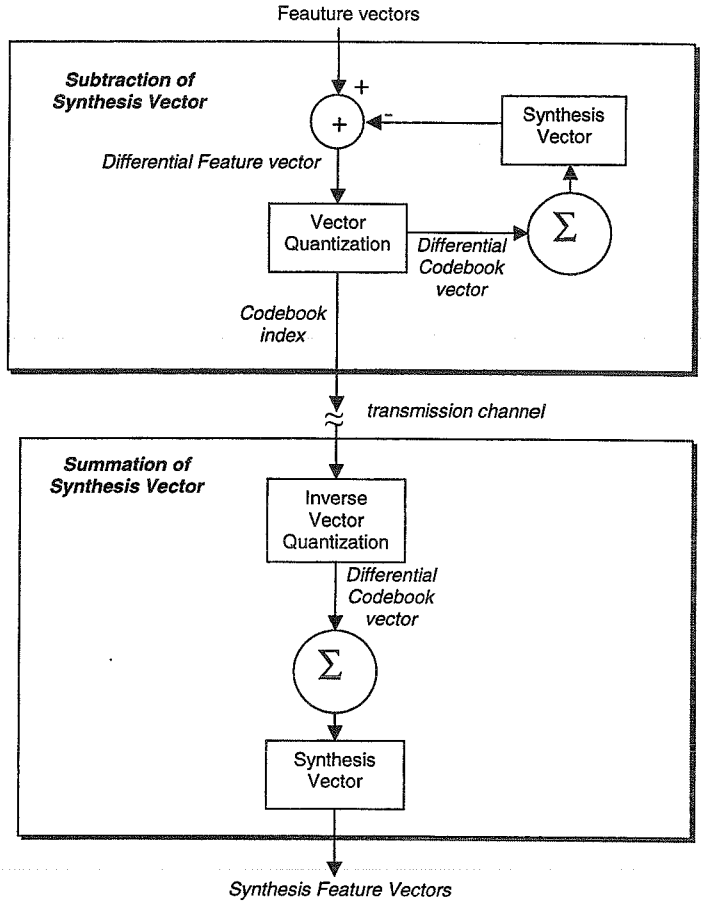

Fig. 2: Differential Feature Generation and Inverse Differential Generation

3. RESULTS

A speech database of 40 speakers was used for training. The speech signals were sampled at 11 kHz. For the proposed methods, the codebooks were trained by the minimum searching method of the Linde-Buzo-Gray algorithm (Linde, Buzo & Gray, 1980). The bit allocation of our speech codec is listed in Table. 1. Bit allocation is shown for 1.6 kbps.

Parameters	Bits
Differential spectrum magnitude feature	9
Number of periods per frame	3
Pitch Period	7
Total bits /23 ms	19

Table. 1: Bit allocation

In Fig. 3 the original set of time samples was compared to the reconstructed samples.

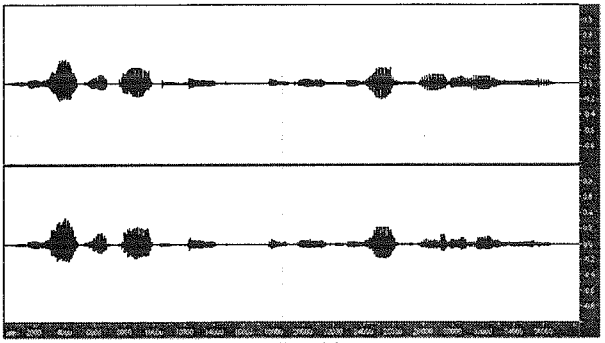

Fig. 3: Speech waveform of the original signal (upper) and after processing with the proposed method (lower diagram)

The proposed method is as the function of the low-rate harmonic coders used to synthesize speech, no phase information is transmitted, which results in a little loss of time alignment between the original speech and the synthesized speech in Fig. 3. This little loss of time alignment makes it not difficult for the proposed coder to perform waveform matching.

Fig. 4 shows the spectrogram of a speech sentence from speech waveforms in Fig. 3. The original spectrogram is shown in the upper image recorded at 11 kHz sampling rate. The spectrogram of the speech signal obtained by our proposed algorithm is depicted in the lower part of Fig. 4. Dark gray areas correspond to the speech components. The light textured area represents background. We can observe from Figs. 3 and 4 that the proposed method achieves a reasonable reconstruction of speech.

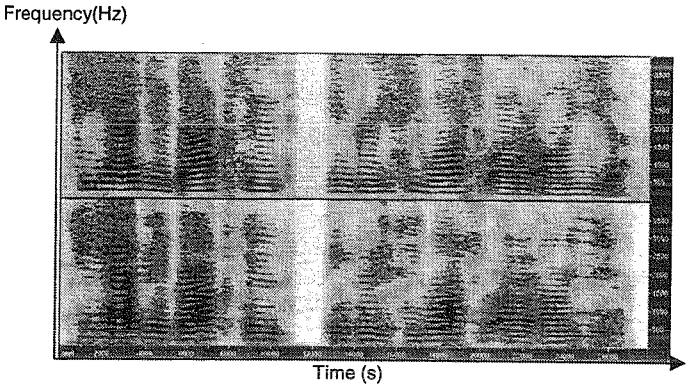

Fig. 2: Speech spectrogram of the original signal (upper) and after processing with the proposed method (lower diagram)

To evaluate the performance of the proposed 1.6 kbps-Codec we used two reference algorithms : a) the 2.4 kbps LPC-10e standard and b) the Harmonic Vector Excitation Coding (HVXC) proposed for MPEG-4 standardization Codecs(Nishiguchi & Matsumoto, 1997). Based on identical speech data the quality of the reconstructed signals obtained from the three methods was judged in an informal mean opinion score (MOS) test by 22 individuals. The results are summarized in Table. 2.

Method	Transmission rate	MOS ± standard deviation
LPC-10e standard	2.4 kbps	2.3 ± 0.35
HVXC	2.0 kbps	3.24 ± 0.16
The proposed Codec	1.6 kbps	3.22 ± 0.25

Table. 2: The proposed algorithm was compared with two reference methods according to an informal mean opinion score (MOS)

REFERENCES

Cho, Y. D., Kim, H. K., Kim, M. Y., S. R. Kim, S. R. (1997) *Pitch Estimation using Spectral Covariance for MBE Vorcoder*, 1997 IEEE Workshop on speech Coding, pp. 21-22, Pocono Manor, PA.

Griffin, D. W. & Jae S. Lim, J. S. (1998) *Multiband Excitation Vocoder* IEEE Trans. On ASSP, Vol. 36, No.8.

Linde. Y., Buzo., Gray. R. M. (1980) *Vector Quantization in Speech Coding*, IEEE Trans. Comm., v28, pp. 84-95.

Lupini, P., Cuperman, V. (1995) *Spectral Excitation Coding of Speech at 2.4 kbitps*, Prpc. ICASSP.

McAulay, R., T. Quatieri, T. (1955) *Sinusoidal Coding" in Speech Coding and Synthesis*, W.B.Kleijn and Paliwal, K. K., Ed., Chapter 4, Elsevier.

M.R.Schroeder, M. R., Atal, B. S. (1985) *Code Excited Linear Prediction (CELP): high quality speech at very low bit rate*, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, IEEE CH2118-8/85, vol. 1, pp. 937-940.

Nishiguchi, M., Matsumoto, J. (1997) *Harmonic Vector Excitation Coding of Speech at 2.0 kbps*, 1997 IEEE Workshop on speech Coding, pp. 21-22, Pocono Manor, PA.