Combined Speech and Audio Coding with Bit Rate and Bandwidth Scalability

Maria Farrugia, Ahmet M. Kondoz

Multimedia Communications Group, CCSR
University of Surrey
Guildford, Surrey
GU2 7XH, UK
M.Farrugia@eim.surrey.ac.uk

Abstract

The growing demand for streaming multimedia services over the Internet and recently also over mobile networks has initiated a great interest in coding algorithms which are able to adapt to different transmission environments and to operate under multiple constraints of bit rate, complexity, delay, robustness to bit errors and diversity of input signals. In the light of these recent developments, we present a novel scalable representation for speech and audio signals with low delay. The algorithm operates in four modes, each based on backward-adaptive linear predictive coding (BA LPC). The first mode is referred to as the base-line narrowband (0–4kHz) coder. Wideband speech and audio signals (0–8kHz) are efficiently represented by the second mode which employs a QMF to split the spectrum into two equal bands. The remaining two modes use a two-stage QMF structure to decompose the bandwidth of 32kHz sampled signals into four bands. Scalability is achieved by means of discrete quantisation layers representing various levels of enhancements for each band and also flexibility in terms of complexity and bit allocation requirements depending on the particular application and on the network resources. The resulting bit rates range from 12 to 64kb/s. The performance of the coder is evaluated by comparing it to MPEG and ITU standards.

1. Introduction

In recent years, there has been a great deal of interest in developing good quality combined speech and audio coding schemes operating at low to medium bit rates. Currently, research in this area is concentrating on scalability, for deployment in fixed wired and mobile communications, such as audio streaming applications. The recently standardised MPEG-4 [1] coder, which comprises a family of algorithms providing speech and audio at rates from 200b/s to 64kb/s, is one example. Scalability increases the end-to-end quality services by avoiding tandem across various networks. Furthermore, scalable coders accommodate varying resource allocations and support prioritisation for packet-based transmissions.

In view of this research trend, this paper presents a scalable coder based on analysis-by-synthesis (AbS) LPC. This powerful technique is widely employed for redundancy reduction in speech. However, it is not commonly used for audio coding due to the fact that music signals are in general not as stationary as speech signals. Instead, audio coders combine efficient frequency-domain compression techniques with psychoacoustic masking in order to achieve high output quality. However, these coders are characterised by high bit rates and large delays and hence are not suitable for deployment in bandwidth restrictive environments, such as wireless communications, and for applications that require low delay. Research in generic speech and audio coding is focusing on target bit rates of 16, 24 and 32kb/s [2] aimed at both fixed and packet transmission applications. The ITU-T has recently been involved in the standardisation of a family of wideband speech and audio coders operating at the same bit rates. The new coding schemes are required to exhibit a similar performance as the G.722 [3] coder at its respective rates under most operating conditions.

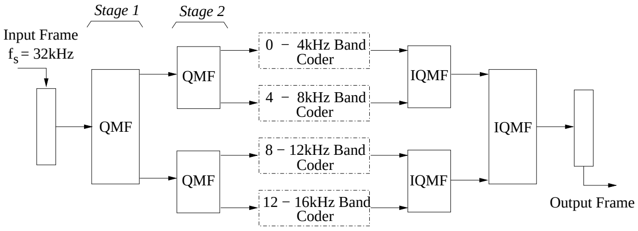

Figure 1: *Schematic of the proposed coder.*

The LPC method is employed for the low bit rate modes of the MPEG-4 standard. However, unlike the MPEG-4 standard, the coder proposed adopts a single approach which is suitable for compressing both speech and audio signals at low bit rates for low cost transmissions as well as at high bit rates for high quality audio on demand. For a better representation of higher bandwidth signals (>4kHz), the proposed algorithm employs QMFs to split the input signal spectrum into narrow bands which are individually encoded, as shown in Figure 1. Spectral modelling is achieved by high order BA LPC analysis over very short frames. The search for the optimum excitation sequence is performed in a closed loop consisting of an excitation generator which produces a number of possible candidates comprising sparse unity magnitude pulses. Besides introducing a number of advantages in terms of storage and search complexity, this structure is suitable for coding highly transitional segments, characteristic of music signals. The available bit rate is allocated among the bands depending on their respective perceptual importance. For a higher quality, a number of enhancement layers are embedded in the encoded bit stream at the expense of higher bit rates.

2. Description of the Base-line Coder

The base-line encoder is displayed in Figure 2. The input signal, sampled at 8kHz, is partitioned into frames of 12 samples (1.5ms). Each frame is perceptually weighted by a 10-tap fil-

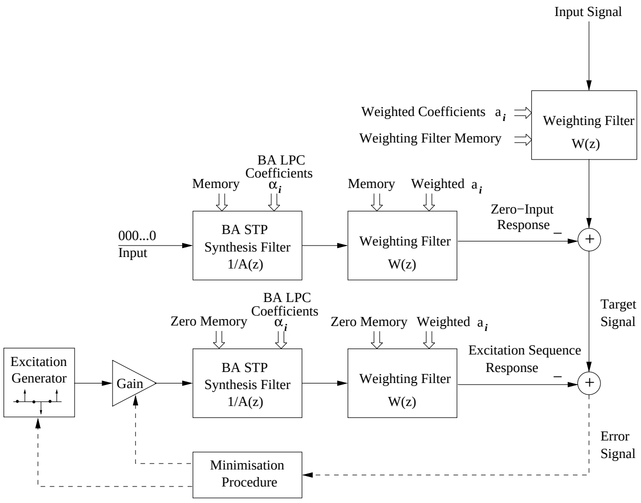

Figure 2: *Base-line encoder.*

ter which aims to provide noise shaping and hence exploit the psychoacoustic effects of human perception. Due to the limited number of bits per frame, a short-term BA LPC filter is employed for spectral modelling. The filter coefficients are calculated using the previously synthesised samples. Hence no extra bits are required for the quantisation of these parameters. A 50-tap filter is chosen in order to cover a time span of at least one complete female pitch period resulting in a reduction in the quality degradation of female speech incurred by the lack of pitch prediction conventionally employed in speech coders.

Table 1: *Potential positions of individual pulses.*

Track	No. of Pulses	Possible Pulse Positions
1	2	0, 3, 6, 9
2	2	1, 4, 7, 10
3	2	2, 5, 8, 11

Referring to the diagram in Figure 2, a target signal is obtained by removing the contribution of the combined LPC synthesis filter and weighting filter memories from the perceptually weighted input signal. Similarly to the G.728 [4] coder, a 10-tap BA gain predictor is employed in order to reduce the dynamic range of the excitation gain and hence achieve a greater accuracy during quantisation. The algorithm then searches for the optimum shape and gain to represent the excitation sequence at the decoder. Similarly to the GSM-EFR 12.2kb/s [5], this is achieved by a pulse-excited model. In order to improve the quantisation performance, the search for these two parameters is performed in a closed-loop fashion. The popularity of this structure stems from the fact that it best fulfils the performance requirements in terms of subjective quality, complexity, robustness and delay. The objective is to generate sets of vectors consisting of sparse unity magnitude pulses placed in predetermined locations. The scheme employed in the base-line encoder, which is displayed in Table 1, allows a uniform representation of all the possible pulse locations. The 12 possible positions in the excitation vector are divided into 3 interleaved tracks, each track containing two pulses of magnitude ±1. Each candidate vector is fed through the cascaded LPC synthesis filter and perceptual weighting filter starting from an all-zero initial state. The sum of the squared error between each synthesised signal and the target signal is computed. Out of the total number of possibilities, the search identifies the best set of pulses which result in minimal distortion. The selected sequence represents the optimal excitation which is transmitted to the decoder in the form of an 18-bit index, as shown in Table 4.

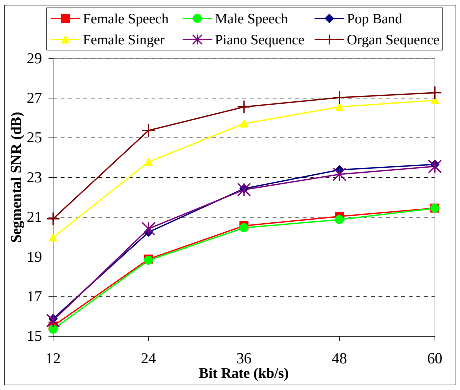

Figure 3: SNR_{SEG} *for additional quantisation layers.*

Finally, the states of the LPC filter and the weighting filter are updated in preparation for the computation of the target signal for the next frame. The filter parameters are updated every 3ms. This interval is short enough for the signal parameters to be assumed approximately constant.

2.1. Multi-Stage Scalable Quantisation

In [6], it was shown that a gradual improvement in the output quality can be achieved by employing multi-layered quantisation at the expense of an increased bit rate. The optimal excitation vector from the first stage of quantisation is synthesised, weighted and subtracted from the target signal to produce a new target signal with lower short-term power. The goal of the closed-loop search in the second stage is to select the excitation vector which when synthesised, matches best the new target signal. At the decoder end, the two decoded excitation vectors are scaled with their respective gain values, added together and fed through the synthesis filter to form the reconstructed output.

The bit stream is embedded, with each subsequent layer of quantisation built on the innermost layer. The resulting bit rate increases by integer multiples of 12kb/s. Figure 3 displays the segmental SNR (SNR_{SEG}) as a function of multi-layered quantisation. The transmitted excitation sequences from second and subsequent stages can be deleted in order of significance to achieve progressive bit rate reductions. The minimum transmission rate is determined by the excitation of the base-line coder, which has the most significance. Hence the scalability feature is an effective means of controlling channel congestion without recoding the input signal.

When the upper layers are discarded, the decoder is still required to produce a meaningful output. Since the encoder has no information on which parts of the bit stream have reached the decoder, the memories of the LP filters are updated by the excitation vector of the first quantisation stage only. Experiments show that for a second stage of quantisation, the SNR_{SEG} is reduced by around 1.5dB compared to when the memories are updated with every stage. However, this penalty is acceptable when considering the several advantages of scalability.

3. Coding Wider Bandwidth Signals

The base-line coder was tested for speech and audio signals sampled at 16kHz and 32kHz. Informal listening tests results showed that the output suffers from noise located in the upper frequencies. This degradation is due to the large dynamic variation between the low and high frequencies which cannot be accurately modelled by the LPC analysis. This problem is overcome by employing 24-tap QMFs which split the input spectrum into flatter 4kHz bands. Each band is thereafter separately encoded. Equal spectral decomposition provides a simple and efficient integer band sampling method. In addition, the spectrum of speech signals can be essentially regarded as consisting of a base band (50–4000Hz) containing most of the perceptually important information and an upper band contributing to the overall perceived intelligibility. Furthermore, equal bands result in a more uniform structure suitable for both bit rate and bandwidth scalability.

The coder is designed to operate in three different modes: Mode 1–3. Mode 1 consists of the base-line coder which has been described in the previous section and which is suitable for telephone bandwidth signals. In Mode 2, the spectrum is decomposed in two bands. This mode is suitable for wideband signals. In Mode 3, the spectrum is split into four bands by a two-stage tree-structure QMF. This mode is suitable for music signals sampled at 32kHz. The algorithmic delay is a function of the sampling frequency. Its maximum value is 165 samples.

Although the upper spectral bands (>4kHz) do not contain as much information as the lower band, there is still a significant amount of correlation present which can be exploited during the encoding process. The design proposed maintains the AbS BA LPC structure employed in the lower band. However, following investigations carried out with different input signals, a 30-tap filter is found adequate to remove the short-term correlation present in the upper bands. Another modification is performed on the dimension of the excitation vector. For Mode 2 and Mode 3, the size of the innovation is set to 24 samples. The pulse excitation structures chosen for the upper three bands are displayed in Tables 2 and 3.

Table 2: *Pulse positions for the 4–8kHz band.*

Track	No. of Pulses	Possible Pulse Positions
1	5	$0, \ldots, 23$

Table 3: *8–12kHz and 12–16kHz bands: Excitation structure.*

Track	No. of Pulses	Possible Pulse Positions
1	1	0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22
2	1	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23

3.1. Bit Allocation

Spectral decomposition allows unequal distribution of the available bit rate among the bands. This approach has been adopted following investigations into the auditory system which have shown that the human ear is most sensitive to low frequencies (up to 4kHz). Hence, the lower band, being perceptually more important, is encoded using a greater percentage of the bit rate.

Both adaptive and fixed bit allocations were considered. In adaptive bit allocation, the quantisation scheme employed in each of the bands is varied every 3ms depending on the spectral energy of the input signal. However, the output bit rate is maintained constant. Such frequent analysis is adequate to trace any rapid transitions in the signal. In addition, the calculation of the RMS value representing the energy content, does not require any lookahead and hence no extra delay is incurred. The first bit allocation scheme considered is displayed in Table 4. The second scheme reduces the bit rate of the lower band to 10.67kb/s. The remaining bits are assigned to the 4–8kHz band. Subjective tests have shown a slight increase in the performance for music signals, as a result of adaptive bit allocation. However, this enhancement is too small to justify the extra computational complexity and the additional bits required to represent the selected quantisation mode. Furthermore, this approach is not very suitable for short frame sizes due to the limited number of bits available for the quantisation of the residual signal. Hence a fixed bit allocation is adopted as shown in Table 4.

Table 4: *Bit allocation for each band.*

Parameters	0–4kHz	4–8kHz	8–12kHz	12–16kHz
Pulse Positions	12	16	8	8
Pulse Polarities	3	5	2	2
Excitation Gain	3	3	2	2
Bit Rate	12kb/s	8kb/s	4kb/s	4kb/s

3.2. Bandwidth and Bit Rate Scalability

Table 6 displays multi-layered quantisation for each of the four spectral bands. The upper bit rate is limited by the target application. Other bit allocation schemes are possible. However, informal listening tests have shown a preference for more accurate quantisation of the lower bands compared to compressed sequences with a wider bandwidth but with coarser quantisation of each band.

4. Channel Error Robustness

In [7], we have shown that the approach proposed is suitable for streaming audio services which make use of the link-layer retransmission schemes offered by EGPRS. Spectral decomposition allows a multi-level retransmission prioritisation scheme which exploits the difference in the perceptual importance of the bands encoded at different bit rates. By allocating separate retransmission criteria for the separate categories of audio packets, significant improvement in channel error robustness is achieved.

5. Performance Evaluation

The subjective quality of the coder proposed is evaluated for different bandwidths and bit rates using the MOS scale. Listening tests were carried out using 14 different sequences consisting of female and male speech sequences, vocal music sequences and instrumental files. 16 subjects took part in the test using a binaural headset.

Table 5 shows the performance of the coder for the chosen signals sampled at 8kHz and coded at 12kb/s. The G.728 standard operating at 16kb/s and the GSM-EFR coder operating at 12.2kb/s are chosen as a reference. The results indicate that the quality obtained at 12kb/s is slightly better than the quality produced by the G.728 standard, for both speech and audio signals. When compared with the GSM-EFR coder at approximately the same bit rate, the proposed coder achieves slightly lower MOS for speech signals. However, the audio signals encoded with the proposed coder are preferred to those produced by the GSM-EFR coding algorithm. The outcome of this test clearly shows that at 12kb/s, the base-line coder is suitable for compressing both speech and audio signals with good quality.

Table 5: *MOS test results for the base-line coder.*

Test Material	Base-line Coder (12kb/s)	G.728 (16kb/s)	GSM-EFR (12.2kb/s)
Female Speech	3.4	3.1	3.7
Male Speech	3.6	3.6	4.2
Female Singer	3.6	3.2	3.4
Male Singer	3.5	3.2	2.9
Piano	3.8	3.6	3.0
Violin	4.2	3.3	3.5
Trumpet	3.9	3.6	3.7

Further tests are carried out to assess the performance of the proposed wideband coding scheme at 20kb/s by comparing it to the quality of the G.722 coder at 48kb/s. Table 7 shows the scoring for the chosen test material. The indicative tests show that the speech quality of the G.722 standard is slightly better than that obtained by the proposed coder at around three times the bit rate. However, the perceptual quality achieved for the male and female singing and the instrumental music sequences is comparable to the quality of the reference coder.

Finally, the performance of the proposed coder for signals sampled at 32kHz and coded at 28kb/s is compared to the quality produced by the MPEG-1 Layer III coder operating at 32, 48 and 64kb/s. The results of this test are shown in Table 8. The scores indicate that at 28kb/s the proposed coder achieves better quality than the MPEG-1 Layer III at 32kb/s and comparable quality to that achieved by the standard at 48kb/s.

6. Conclusions

In this paper, we have proposed a combined approach for compressing speech and audio signals sampled at 8, 16 and 32kHz. The coder is based on a BA LPC analysis for spectral modelling in conjunction with a pulse excitation structure for efficient representation of the residual signal. The perceptual imbalance that exists between the lower and upper frequencies is exploited by employing spectral decomposition together with different resolution of quantisation and unequal error protection for each band. Other key features include low delay and scalability of the signal bandwidth, output throughput and implementation complexity. Scalability is achieved by multi-layered embedded quantisation. This property is attractive for optimal use of bandwidth-varying transmission channels as currently found in the Internet and expected to be provided by third generation mobile systems such as UMTS. In addition, listening test results showed that the coder achieved acceptable quality over a diverse class of input signals, comparable to ITU and MPEG standards at similar bit rates.

Table 7: *MOS test results for the wideband coder.*

Test Material	Wideband Coder (20kb/s)	G.722 Coder (48kb/s)
Female Speech	3.7	4.2
Male Speech	4.0	4.3
Female Singer	3.6	3.6
Male Singer	3.2	3.4
Violin	3.7	2.6
Piano	3.7	3.6
Castanets	4.2	3.9

Table 8: *MOS test results for 32kHz sampled signals.*

Test Material	Coder	MPEG-1 Layer III		
	(28kb/s)	(32kb/s)	(48kb/s)	(64kb/s)
Female Singer	3.6	2.7	4.4	4.6
Male Singer	4.3	3.7	4.4	4.4
Violin	4.2	3.8	4.4	4.6
Piano	4.1	3.9	4.4	4.7
Castanets	4.4	4.0	4.1	4.7

7. References

[1] Painter, T. and Spanias, A., “ Perceptual Coding of Digital Audio”, Proc. of the IEEE, p 449–708, April 2000.

[2] Combescure, P. and et al, “A 16, 24, 32 kbit/s Wideband Speech Codec based on ATCELP”, ICASSP ’99, Arizona.

[3] CCITT Rec. G.722, “7 kHz Audio-Coding within 64 kbit/s”, Melbourne, 1988.

[4] CCITT Rec. G.728, “Coding of Speech at 16kbit/s using low-delay Code Excited Linear prediction”, September 1992.

[5] ETSI/TC SMG, “Rec. GSM 06.60: Enhanced Full Rate Speech Transcoding”, V4.1.0, June 1998.

[6] Farrugia, M., Jbira, A., and A.M. Kondoz, “Pulse-Excited LPC for Wideband Speech and Audio Coding”, IEE Electronics & Communications Colloquium: Audio and Music Technology: The Challenge of Creative DSP, London, November 1998.

[7] Fabri, S., Farrugia, M. and Kondoz, A.M., ”Provision of Scalable Streaming Audio Services over GPRS”. To appear in *IEEE VTC*, Greece, May 2001.

Table 6: *Layers of quantisation for various bit rates and bandwidths.*

Spectral Band	Bit Rates (kb/s)												
	12	20	24	28	32	36	40	44	48	52	56	60	64
0–4kHz	1	1	2	1	2	3	2	3	2	3	2	3	4
4–8kHz	0	1	0	1	1	0	2	1	2	2	2	1	2
8–12kHz	0	0	0	1	0	0	0	0	1	0	2	1	0
12–16kHz	0	0	0	1	0	0	0	0	1	0	2	0	0