HYBRID MULTI-MODE/MULTI-RATE CS-ACELP SPEECH CODING FOR ADAPTIVE VOICE OVER IP

G. Ruggeri, F. Beritelli, S. Casale

Dipartimento di Ingegneria Informatica e delle Telecomunicazioni - University of Catania
Viale Andrea Doria 6 95125 Catania
email: (beritelli, gruggeri, scasale)@iit.unict.it

ABSTRACT

This paper presents a hybrid Multi-Mode/Multi-Rate, toll quality CS-ACELP coder developed for Voice over IP applications. The coder uses coding modes compatible with the three 6.4, 8, and 11.8 kbit/s coding schemes standardised by ITU-T in G.729. In particular, the algorithm presents 4 coding categories, with an average bit rate ranging between about 3 and 8 kbit/s, that adapt the rate to changes in network conditions.

1. INTRODUCTION

In the last few years an increasing amount of attention has been paid to technologies for the transmission of voice over data networks. However if voice transmission over IP (Internet Protocol) networks is to be competitive with transmission on PSTN (Public Switched Telephone Networks), users will have to be provided with a quality of service (QoS) that is comparable to that of circuit switching networks. This paper proposes a hybrid speech coding solution for Adaptive Voice over IP. In particular we propose an intrastandard, Multi-Modo/Multi-Rate (M^3R), toll quality speech coder based on the structure of the CS-ACELP algorithm of G.729 at 8kbit/s [1]. The hybrid coder integrates the two extensions to G.729 at 6.4 kbit/s [2] and 11.8 kbit/s [3] recently standardised by ITU-T, and also exploits new fuzzy pattern recognition techniques for multimode classification and new coding models for the different phonetic classes considered.

2. THE G.729 CODER AND ITS EXTENTIONS

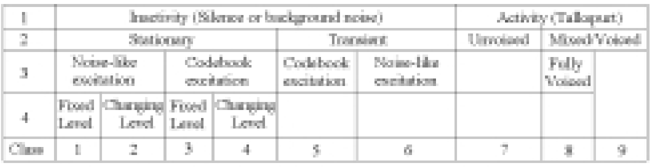
G.729 at 8kbit/s is a recent good quality ITU-T speech coding standard based on conjugate-structure algebraic-code-excited linear prediction (CS-ACELP) [1]. More recently, ITU-T has standardised two extensions of the 6.4 kbit/s and 11.8 kbit/s G. 729, respectively indicated as G.729 Annex D [2] and Annex E [3]. The coding architecture of the former, which is a low-bit-rate extension of Recommendation G.729, is identical to that of G.729. There are, however, certain differences, i.e. the use of a reduced codebook and less fine quantisation of some parameters such as pitch delay and gain. The latter is a higher-bit-rate extension of G.729. Here again the basic scheme is the same as that of G.729 but there are some variations that do not only concern the quantisation of the parameters. The most important novelty is the introduction of backward linear prediction analysis, for better coding of music and speech uttered in the presence of stationary noise. The bit allocation for G.729 and its extension are summarised in Table 1. The two extensions to G.729 have been a significant reference point for the development of the hybrid multimode/multirate coder presented in the following sections.

Parameters	Annex D	G.729	Annex E Forward	Annex E Backward
LPC	18	18	18	
Pitch period	12	13	13	13
Parity bit	0	1	1+1+1	1+1+1
codebook	22	34	70	88
Pitch and codebook gains	12	14	14	14
TOTAL	64	80	118	118

Table 1. Bit allocation, every 10ms, for G.729 and other operative modes.

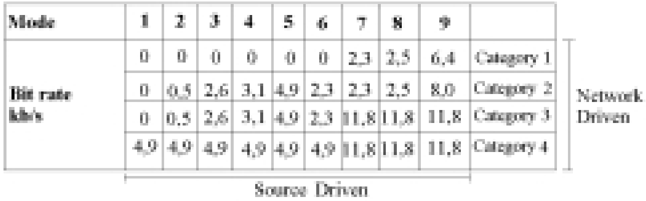
3. HYBRID CODING PROPOSED

To guarantee a certain QoS even in critical conditions featuring great delays and background noise, it is necessary to control the peak rate, and therefore use a multirate codec, and also to have good comfort noise models that only multimode coding can provide. The two extensions to G.729, Annex D and Annex E, recently standardised by ITU-T, and a previous work on a robust multimode coder based on the G.729 structure [4], have been the starting point for the development of a hybrid Multi-Modo/Multi-Rate (M^3R) codec. The new codec guarantees a toll quality, it is



1	Inactivity (Silence or background noise)						Activity (Talkspurt)		
2	Stationary				Transient		Unvoiced	Mixed/Voiced	
3	Noise-like excitation		Codebook excitation		Codebook excitation	Noise-like excitation		Fully Voiced	
4	Fixed Level	Changing Level	Fixed Level	Changing Level					
Class	1	2	3	4	5	6	7	8	9

Fig. 1. Phonetic classes considered



Mode	1	2	3	4	5	6	7	8	9	
Bit rate kb/s	0	0	0	0	0	0	2,3	2,5	6,4	Category 1
	0	0,5	2,6	3,1	4,9	2,3	2,3	2,5	8,0	Category 2
	0	0.5	2,6	3,1	4,9	2,3	11,8	11,8	11,8	Category 3
	4,9	4,9	4,9	4,9	4,9	4,9	11,8	11,8	11,8	Category 4

Network Driven

Source Driven

Fig. 2. Coding modes and categories

robust to background noise, and also has the advantage of being intrastandard, i.e. it fully exploits the architecture of G.729, while maintaining compatibility with both the basic version at 8 kbit/s and the two extesnions at 6.4 kbit/s and 11.8 kbit/s. The new features introduced are mainly the insertion of a phonetic classifier that is robust to background noise, on which the multimode part of the codec is based; the addition of the two modes compatible with annex D and E of G.729, with which the coder's bit rate is adapted, and the insertion of new coding models for fully voiced sounds and the synthesis of background noise.

3.1. Robust phonetic classification

Phonetic classification is undoubtedly one of the most delicate issues in multimode speech coding. The coding of a speech segment by means of an inappropriate model, in fact, causes a degradation in quality, which in some cases reaches unacceptable values. In order to have a speech source driven coder, it needs to exploit fully the large amount of pauses during a conversation and the great variations in the characteristics both of active speech and background noise. The classifier architecture proposed presents four levels of classification characterised by the 9 phonetic classes shown in Figure 1. More specifically, at the first level a Voice Activity Detector (VAD) distinguishes activity segments (talkspurts) from non-activity segments (silence or background noise). At the second level of classification, if the speech segment is active, a voicing detection algorithm discriminates between unvoiced (UV) and voiced (V) sounds. The VAD and V/UV detectors are based on an adaptive version of the algorithms recently proposed in [5] and [6] respectively, where it is demonstrated that they guarantee a greater robustness to background noise than traditional solutions. In the category of voiced sounds the fully voiced speech segments are identified by a backward cross-correlation algorithm described in [4]. Every frame classified as inactive (silence or background noise) is initially separated into stationary and non-stationary. Both stationary and non-stationary frames are closed-loop classified as acoustic noise segments requiring a noise-like or codebook LPC excitation. Finally, within stationary frames, by means of a simple change in level control, the speech segments that maintain their energy level close to that of the previous frame are distinguished from those that have undergone a variation.

Pulse	Sign	Position
I_0	+1	0,5,10,15,20,25,30,35
I_1	-1	1,6,11,16,21,26,31,36
I_2	+1	2,7,12,17,22,27,32,37
I_3	+1 -1	3,8,13,18,23,28,33,38 4,9,14,19,24,29,34,39

Table 2. Codebook structure used

Noise	Excitation	LPC	Gain
Bus	Codebook	Stationary	Mixed
Car	Noise-like	Stationary	Mixed
Train	Noise-like	Stationary	Mixed
Dump	Codebook	Stationary	Mixed
Cons	Codebook	Stationary	Mixed
Factory Restaurant	Codebook	Mixed	Non Stationary
Street	Codebook	Mixed	Non Stationary
Office Shop	Codebook	Mixed	Non Stationary
Pool Trains	Codebook	Mixed	Non Stationary

Table 3. Noises characterization

3.2. Talkspurts and background noise coding

In order to develop an intrastandard speech coder we used, as the core scheme for the proposed M^3R coder, the structure and the main procedures of the G.729 CS-ACELP algorithm. In particular, we used the standard ITU-T G.729 at 6.4, 8 and 11.8 kbit/s as the coding algorithm for mixed or voiced frames (class 9), whereas for unvoiced sounds we used a simple, time-varying, white Gaussian noise-excited LPC filter (class 7). The noise generator is the same as the one used for the comfort noise system in ITU-T G.729 Annex B. For excitation gain coding we used a non-uniform quantizer with 32 levels, whereas for the other parameters the quantization process was based on the procedures of the G.729 standard. In order to exploit the periodic nature in the steady-state portion of a voiced segment, we used a recent new algorithm for efficient coding of fully voiced sounds at 2.5 kbit/s (class 8) [4]. The method uses a backward cross-correlation measure between the current LPC synthetic residual and the two previous ones. Using a simple coding scheme based on an excited LPC filter, we carried out several subjective tests, coding each type of noise varying the type of excitation and adapting or not adapting the excitation gain and LPC parameters. The codebook structure considered is the same as the one used by the G.729 standard, the only difference being the sign of each pulse, which was fixed a priori according to the position, as shown in Table 2. Although this choice assures good quality in the reconstruction of noise, the bit rate is kept below 5 kbit/s. The results of the tests are given in Table 3. For a toll quality reconstruction of noise, some types of background noise require a noise-like residual whereas others require codebook ex-

Mode	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
LPC	-	-	-	-	18	18	18	18	Look at Table 1
Gain	-	5	-	5	5	5	5	-	
Codebok	-	-	26	26	26	-	-	-	
Pitch	-	-	-	-	-	-	-	-	
Exitation Lag	-	-	-	-	-	-	-	7	
Parity	-	-	-	-	-	-	-	-	
Bit/frame	**0**	**5**	**26**	**31**	**49**	**23**	**23**	**25**	

Table 4. Parameters trasmitted and bit allocation

citation. In addition, for some types of noise it is not necessary to update the LPC parameters, and for others again not even the excitation gain. In these cases the bit rate is reduced even further. For non-stationary background noise coding we used the unvoiced speech coding model for acoustic noise with noise-like residual (class 6) and a time-varying LPC filter excited by a codebook for acoustic noise characterized by an LPC residual with a non-flat spectrum (class 5). Choice of the type of excitation is made on the basis of a threshold comparison of the ratio between the Weighted Mean Square Errors (WMSE) calculated in the two cases of Gaussian and codebook excitation. For stationary background noise we developed four coding models based on both a similar control performed on the flatness of the LPC residual spectrum, and the changing level of the LPC residual, by means of a simple threshold comparison. Table 4 shows the parameters transmitted and the bit allocation for each mode. For the codec to be network-driven as well, we grouped the coding modes in the four categories illustrated in Fig.2 so as to have different average bit rates according to the load on the network. The coding category 2, for average workloads, uses the 8 kbit/s G. 729 as the peak rate. In a situation of network congestion, optimal talkspurt and background noise is forgone in order to minimise the throughput due to the coder. In this case (category 1) no distinction is made between the classes of background noise: they are all reconstructed using a traditional comfort noise system. Active frames are coded as coding category 2 except for mixed sounds, which are coded at 6,4 kb/s as established in G.729 Annex D. Category 3, envisaged for use in low network load conditions, has a higher bit rate as it uses the 11.8 kbit/s extension to G. 729 for activity periods. Finally, category 4, the one with the highest rate, is envisaged for use in very low network load conditions. When the coder operates in this category it selects annex E at 11,8 kb/s for talkspurts, while all frames classified as noise are coded at 4.9 kbit/s using mode 5.

4. COMPARISON WITH $G.729/DTX$

We considered the 4 operating categories and compared them with G.729 at 8kb/s in the discontinuous transmission mode (DTX) specified in G.729 Annex B [7]. The performance of each mode was evaluated in terms of average bit-rate and perceptive quality. For each mode we carried out a series of tests considering several acoustic noise environments and SNR levels. The speech database used consists of several speech sequences, sampled at 8 kHz, linearly quantized at 16 bits and levelled at 26 dB below codec overload. The sequences, spoken in Italian by male and female speakers, each last 6 minutes with an activity factor of 40 %. Table 5 shows the percentage of phonetic classes selection while Table 6 shows the average bit-rate, with varying SNRs (0, 10, 20 dB) and types of additive background noise (car, traffic, babble), besides the clean case. The results show how well the phonetic classifier works: for example, the speech coding models developed for background noise with a varying LPC residual spectrum (class 3, 4, and 5) are rarely selected for conversations in the presence of a stationary signal like car noise, whereas with traffic or babble noise they are selected frequently. The opposite happens when classes 1 and 2 are selected. In addition, in noisy environments the percentage of selection of class 6 is considerably reduced, as sounds with a noise-like residual, such as unvoiced sounds, are transformed into sounds requiring a codebook excitation. Naturally, as the average bit rate is closely linked to the behaviour of the phonetic classifier, it depends on both the nature of the background noise and the SNR. The lower bit-rate cases occur in quiet acoustic conditions, as class 1 is chosen more frequently. In the clean case, for example, the average bit-rate is 2.2 kbit/s, for the coding category 2 as class 1 is selected in about 55 % of cases, which corresponds to the percentage of silence frames present in the conversation. The worst case refers to car noise at SNR=0 dB in that, due to the high noise-to-talkspurt misclassifications introduced by the VAD, 71% of the time the codec selects class 9 at the higher rate. A series of informal listening tests based on the Comparison Category Rating (CCR) method [8] were carried out by 24 listeners to evaluate the perceptive quality of the new speech coder. The last row in Table 6 shows the results in terms of Comparison Mean Opinion Score (CMOS) values, i.e. the differences in MOS scores between the coders proposed and the 8 kbit/s G.729 standard with VAD Annex B. We have on average a degradation of 0.1 MOS scores for coding category 1 due to the use of the low bit-rate extension to G.729. Using coding category 2 on average we have an improvement of about 0.1 MOS due to the use of G.729 at 8 kbit/s, the limits of the VAD and the comfort noise system in G.729. Further, with respect to the simple comfort noise system of G.729, the M^3R coder reconstructs the active speech frames detected as noise by means of the most appropriate coding scheme chosen from the six classes developed for background noise. It should be noted that this comparison is a conservative one for the M^3R codec in that there is also a 20 % reduction in bandwidth as compared with the simple ON-OFF coding of G.729 with VAD. Using the high bit-rate extension of G.729 at 11.8 kbit/s (coding category 3) for talkspurt coding the quality improves on average by 0.3 MOS at the expense of a 20% increase in the bit rate. Finally, for coding

Noise	SNR (dB)	1	2	3	4	5	6	7	8	9
Clean		55.08	3.61	0	0	0	3.22	7.79	8.6	21.7
Car	00	13.49	9.6	0.26	0.31	0.12	0.79	0.34	3.76	71.19
	10	21.72	7.12	0	0	0	0.98	1.83	6.57	61.53
	20	39.64	10.8	0	0	0	1.9	4.8	8.23	34.55
Traffic	00	12.23	10.3	13.6	10.51	12.43	4.18	1.09	2.13	33.47
	10	18.98	13.67	7.11	5.18	9.85	5.53	2.79	4.31	32.55
	20	38.87	11.91	0.15	0.16	0.46	8.43	3.9	7.9	28.09
Babble	00	2.36	1.69	20	15.66	23.04	2.28	1.03	1.15	32.76
	10	2.18	1.52	17.06	13.23	19.73	2.08	2.34	3.9	37.9
	20	15.83	11.07	2.75	2.9	4.26	16.53	4.97	6.03	35.55
Average		22.04	8.13	6.09	4.79	6.98	4.59	3.09	5.26	38.93

Table 5. Percentage of phonetic classes selection in varying acoustic condition

Noise	SNR (dB)	Average rate for M^3R in category 4	Average rate for M^3R in category 3	Average rate for M^3R in category 2	Average rate for M^3R in category 1	Average rate G729/DTX
Clean		7.58	4.58	2.22	1.78	2.52
Car	00	10.08	8.97	5.88	4.65	7.15
	10	9.71	8.30	5.19	4.14	7.36
	20	8.17	5.71	3.18	2.53	6.05
Traffic	00	7.42	5.76	4.19	2.22	5.07
	10	7.63	5.70	3.80	2.25	5.12
	20	7.64	4.99	2.81	2.08	4.69
Babble	00	7.30	6.31	4.86	2.14	4.69
	10	7.94	7.08	5.05	2.57	4.78
	20	8.10	6.29	3.91	2.54	5.18
Average		8.15	6.37	4.11	2.69	5.31
CMOS versus G729/DTX		0.5	0.3	0.1	-0.1	

Table 6. Average bit rate and CMOS

category 4 there is an improvement of 0.5 MOS as for all types of background noise we use the 4.9 kbit/s codebook excitation mode. In this case there is an increase in bandwidth of about 60%.

5. CONCLUSIONS

Starting with the CS-ACELP architecture of the G.729 standard at 8 kbit/s and considering the two extensions at 6.4 and 11.8 kbit/s recently standardised by ITU-T as annex D and E to G.729, we have proposed a hybrid Multi-Mode/Multi-Rate intrastandard codec. It can be both source-driven in 9 phonetic classes, and network-driven in 4 coding categories so as to select different average bit rates ranging from about 3 kbit/s to 8 kbit/s. If the codec is used together with a control mechanism, it is capable in any network conditions of finding the right trade-off between the two main factors that determine quality of service.

6. REFERENCES

[1] Rec. G.729, "Coding of speech at 8 kbit/s using conuigate-structure algebraic-codebook-excited linear prediction (cs-acelp)," Tech. Rep., ITU-T, Feb. 1996.

[2] Rec. G.729 Annex D, "Annex d to recommendation g.729: 6.4kbit/s cs-acelp speech coding," Tech. Rep., ITU-T, Sept. 1998.

[3] Rec. G.729 Annex E, "Annex d to recommendation g.729: 6.4kbit/s cs-acelp speech coding," Tech. Rep., ITU-T, Sept. 1998.

[4] F. Beritelli, "A modified cs-acelp algorithm for variable rate speech coding robust in noisy environments," *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 31–34, February 1999.

[5] F. Beritelli, S. Casale, and A. Cavallaro, "A fuzzy logic-based speech detection algorithm for communications in noisy environments," *Proc. of IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP' 98)*, pp. 565–568, Seattle, USA, 12-15 May 1998.

[6] F. Beritelli and S. Casale, "Robust voiced/unvoiced speech classification using fuzzy rules," *Proc. Of IEEE Workshop on Speech Coding*, pp. 5–6, Pocono Manor, Pensylvania, USA, 7-10 September 1997.

[7] Rec. G.729 Annex B, "A silence compression scheme for g.729 optimized for terminal conforming," Tech. Rep., ITU-T, Oct. 1996.

[8] Rec. P.800, "Methods for subjective determination of trasmission quality," Tech. Rep., ITU-T, August 1996.