A Study of Variable Pulse Allocation for MPE and CELP Coders Based on PESQ Analysis

Shi-Han Chen[1]*, Kuo-Guan Wu*[2]*, and Chih-Chung Kuo*[1]

[1]Advanced Technology Center, CCL, ITRI, Hsinchu, Taiwan
[2] Dept. Electrical Engineering, National Chung Hsing University, Taichung, Taiwan
korochen@itri.org.tw

10.21437/Interspeech.2005-816

Abstract

A novel scheme of allocating variable pulses for each frame is proposed to reduce the bit-rate of MPE and CELP coders while maintaining the same speech quality. Since speech signal is not stationary, the required pulse number in a speech coder should be variable frame by frame. In this paper we tried to approximate the optimal pulse allocation by greedy search algorithm based on the criterion of perceptual disturbance value derived by PESQ analysis. In the experiments the proposed scheme was used to reduce the pulse numbers of two standard speech coders, G.723.1 and MPEG-4 CELP. The results show that the proposed scheme can achieve over 30% bit-rate reduction in fixed codebook (FCB) and about 20% in all for both coders while maintaining the same speech quality in both objective and subjective measure. We also designed several methods to accelerate the optimal search, which could largely reduce the execution time by 120 times in the best case.

1. Introduction

Analysis-by-Synthesis (AbS) structure [1] is the most successful and commonly used technique in modern speech codecs. In [1] the excitation of each frame is given by a fixed number of pulses, which is known as Multi-Pulse Excited (MPE) codec. However, many different representations of excitation existed. The most well known representations are Regular-Pulse Excited (RPE) codec which is adopted in GSM system, and Code excited linear prediction (CELP) [2] codecs which is included in several important ITU-T standards such as G.723.1 5.3 kbps [3], G.729 (8 kbps), and G.728 (16 kbps). CELP coders can achieve toll-quality encoding of speech signals at bit-rates above 6 kbps. However, at lower bit-rates, due to the shortage of bits for encoding fixed codebook (FCB) excitation, the voice quality of the CELP coder becomes poor [4].

Many previous works were focused on representing excitation more efficiently. RPE and CELP codecs both limit the pulse positions by some simple rules, and they have limited success in reducing the bit-rate of FCB excitation. Some researchers tried to limit the pulse positions by exploiting the energy distribution and the periodicity of FCB pulses [4, 5]. However, in [4] it was mentioned that these properties were only used in encoding voiced and transition frames, and therefore limits the reduction of FCB bit-rate and an additional voice type classifier is needed. Some other researchers tried to take advantage of the time-varying characteristic of speech and used different strategies in encoding different types of speech [6, 7, 8]. Again in these works a robust voice type classifier is always needed.

In this work, we examined the frame disturbance values of a speech coder that can be derived using PESQ and found that the minimum required pulse numbers in different parts of speech to maintain the same speech quality may not be the same. Therefore if we can appropriately allocate the pulse number of each frame according to its frame disturbance value, we can produce an adapted speech with similar speech quality of the standard at a smaller FCB bits. However, since the coding process is not independent across frames, it is hard to derive the "optimal" pulse number solution directly from the frame disturbance values, and exhausted search is not feasible, too. By exploiting the relationship between the disturbance values and PESQ score improvements, we can approximate the optimal pulse allocation by greedy search algorithm based on the criterion of perceptual disturbance value, and by iterative adjustment of the pulse number of each frame, the overall speech quality will eventually reach that of the standard coders at a smaller pulse number. Additionally, several methods were designed and the execution time of the greedy search could be largely reduced. Compared to the former methods [4, 5, 6, 7, 8], the proposed method does not require any voice type information and therefore does not need a voice type classifier. Additionally, the proposed method is shown to be effective for both MPE and CELP based speech coders. Although the complexity of this scheme is larger than the original codecs due to the iteration process, it is acceptable in offline compression of speech and therefore can be used to reduce the footprint of any voice data when they need to be stored on a device with limited storage size, such as PDA or portable music player.

2. Perceptual disturbance analysis of codecs

In this work we use ITU P.862 (PESQ) [9] to perform the perceptual disturbances analysis. For a given source utterance and a corresponding degraded utterance, PESQ calculate a frame disturbance sequence between the two utterances and use these disturbance values to predict the objective PESQ MOS score. The codecs we used in this paper are G.723.1 6.3 kbps, which is a MPE coder, and MPEG-4 CELP 10.3 kbps [10], which is a CELP coder.

Fig. 1 compares the frame disturbance sequence of the G.723.1 6.3 Kbps standard coder with that obtained by changing the FCB pulse number (represented by M) from M=6/5 to M= 3. Since results of MPEG-4 CELP are quite similar, they are not shown here. We can see from the figure that the disturbances obtained by using fewer FCB pulses are quite comparable with those of the standard coder in some parts of speech. Actually, 36% of the disturbance values with M= 3 are smaller than or equal to those of the G.723.1 coder. Roughly speaking, fewer FCB pulses in those frames should be enough to represent the excitation signal and more pulses

are required in those frames with larger disturbance values. This leads to an idea of adaptively allocating the pulse number of a frame according to its disturbance value, and in the next section we will discuss our proposed method.

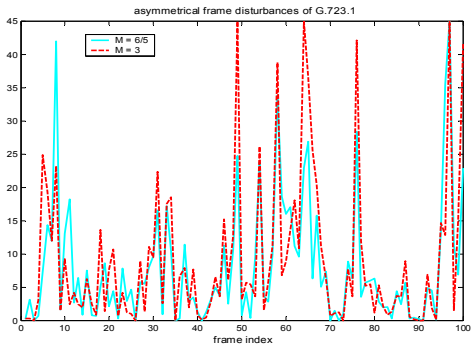

Figure 1: Frame disturbances of the G.723.1 6.3 kbps standard coder and the coder with $M = 3$.

3. The pulse allocation strategy

3.1. Greedy search algorithm based on disturbance values

Fig. 2 depicts the proposed scheme. The adapted speech $s_2(n)$ is produced by transcoding the input speech $s(n)$ using the standard speech coder with a set of pulse number that is "different" from that in the standard. Next the perceptual disturbance values between $s_2(n)$ and $s(n)$ are derived using P.862. Base on the observation described in previous section, if we can appropriately adjust the pulse number of each frame according to its frame disturbance value, we can produce an adapted speech with similar speech quality of the standard at a smaller FCB bits. However, since the coding process is not independent across frames, it is hard to derive the "optimal" solution of pulse number allocation directly from the frame disturbance values, and exhausted search is not feasible, too. Instead we use an iterative adjustment of pulse number and try to find a "sub-optimal" solution.

By observing that the minimum required pulse number of each frame to maintain the same speech quality is not the same, the pulse number is first initialized to the minimum value defined in both speech codecs, which is 1 for G.723.1 and 4 for MPEG-4 CELP. Next in each iteration N frames are chosen to increase their allocated pulse number. In order to decide which N frames are the most appropriate choices, in Fig. 3 we plot the P.862 score improvement when increasing the pulse number of a frame by one v.s. the corresponding frame disturbance value of a G.723.1 transcoded speech. It is clear that when a frame has a larger disturbance value, it tends to have a larger improvement when its pulse number is increased. On the other hand, increasing the pulse number of frames with smaller disturbances may lead to decrease in PESQ score. Therefore the key is to choose N frame with largest frame disturbances to increase their pulse numbers in each iteration. In the next section we will show that we can speed up the iteration process by properly changing the number N, and for convenience we change the notation N to N_{iter} which indicates the number N may change in different iteration. To decide when to stop the iteration process we first produce the PESQ score S_1 of the standard codec transcoded speech. The iteration will stop when score of $s_2(n)$ is equal or larger than S_1, which means its objective speech quality is similar with that of the standard.

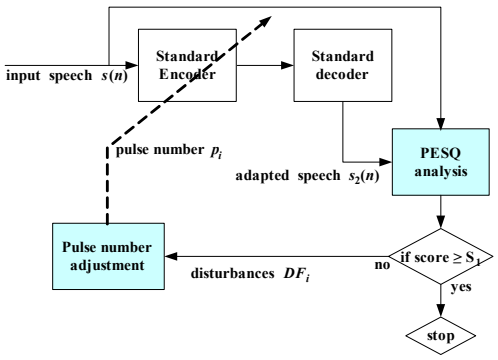

Figure 2: Iterative pulse number allocation mechanism.

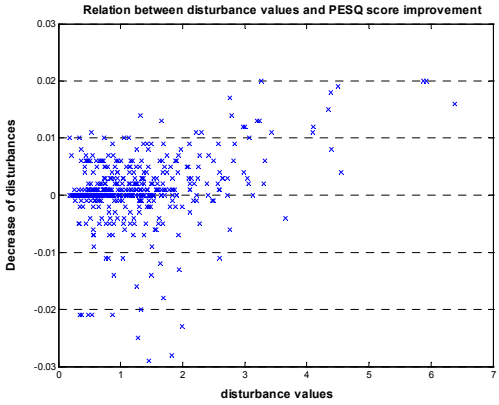

Figure 3: PESQ score improvement of each frame v.s. corresponding frame disturbance value.

3.2. Experiment results

We apply our proposed method in G.723.1 6.3 kbps and MPEG-4 CELP 10.3 kbps and evaluate the performance using sentences from TIDIGITS [11]. The test sentences are randomly selected from the database, and the total length of the chosen sentences is about 1 min for both female and male. In Fig. 4 we apply the proposed method in G.723.1 and we plot the PESQ score of each iteration of a female sentence by setting N_{iter} to a fixed value 1 and 20. As shown in the figure, PESQ score of the adapted speech will eventually reach that of the standard with a much smaller FCB bits. The required FCB bits of different test conditions are shown in Table 1-2 for both male(M) and female(F). 3 additional bits required to store the pulse number for each frame are already included in the FCB bits, and the execution time is represented by multiple of the execution time of the standard. The required FCB bits are smallest when N_{iter} =1 and increase when N_{iter} grows, as shown in Fig. 4. The reason is that when N_{iter} grows, more frames with smaller disturbances will be chosen and increasing the pulse numbers in those frames does not necessary bring score improvement, which is already shown in Fig. 3. However, although N_{iter} =1 has the best performance, the execution time is about 1400 times the original G.723.1, and apparently this is too long for real life application.

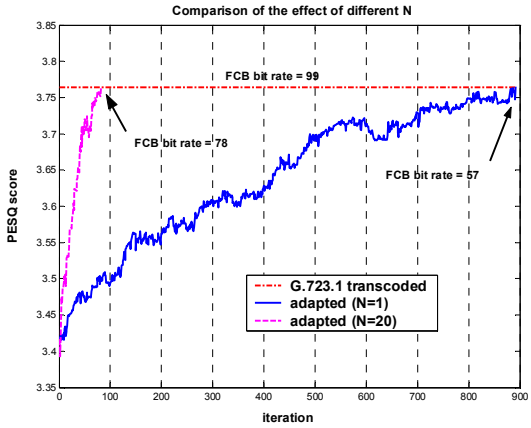

Figure 4: PESQ score and FCB bit rate of the adapted speech v.s. iteration when N_{iter}=1 and N_{iter}=20.

	FCB bits		Execution time	
	F	**M**	**F**	**M**
G.723.1	99	99	1	1
+ max 1	64	60	1411	1295
+ max 20	68	61	82	60
+ max 80	75	66	23	18

Table 1: Results of the case G.723.1.

	FCB bits		Execution time	
	F	**M**	**F**	**M**
MPEG-4 CELP	120	120	1	1
+ max 1	87	75	1787	618
+ max 20	89	76	128	20
+ max 80	91	78	35	10

Table 2: Results of the case MPEG-4 CELP.

4. Enhanced pulse allocation strategy

4.1. Exploiting the relationship between PESQ improvement and frame disturbances

From Fig. 4 we can see that the PESQ score essentially increases along iteration, however, score still decreases in some interval. The reason is that when the disturbance values are decreasing along iteration, frames with smaller disturbances are increasing, and it leads to a higher possibility to increase pulse number in frames with small disturbances. Therefore the most appropriate number of chosen frames should also decrease after some iterations. Next we initialize N_{iter} to 80 and decrease N_{iter} when disturbances increase, since increase in disturbances means some wrong frames are chosen. Results are shown in Table 3-4, and both the FCB bits and execution time are reduced due to the more efficient iteration process.

	FCB bits		Execution time	
	F	**M**	**F**	**M**
G.723.1	99	99	1	1
+ max 80	75	66	23	18
+ max 80decade	68	63	20	17

Table 3: Results of the case G.723.1 with N_{iter} change.

	FCB bits		Execution time	
	F	**M**	**F**	**M**
MPEG-4 CELP	120	120	1	1
+ max 80	91	78	35	10
+ max 80decade	87	75	33	5

Table 4: Results of the case MPEG-4 CELP with N_{iter} change.

4.2. Inter-frame dependence of speech codecs

To further decrease the FCB bits, we try to take advantage of the inter-frame dependence property in speech codecs. Since encoding of a frame depends on the previous encoded and decoded frame, increasing the pulse number of a frame should also decrease the quantization error produced by FCB in that frame and therefore increase the prediction gain in consecutive frames due to the long-term prediction process. In Fig. 5 we plot the statistic of the overall decrease of disturbance in frame $\{1\ldots I\}$ v.s. I when we increase the pulse number by one at frame 1. 6-norm defined in P.862 is used to calculate the overall decrease in disturbance, which is

$$L_6[I] = \left(\frac{1}{I} \sum_{i=1}^{I} disturbance[i]^6 \right)^{1/6} \quad (1)$$

As shown in the figure, this inter-frame dependence in G.723.1 which is a MPE coder is stronger than MPEG-4 CELP which is an ACELP type coder, and clearly the largest improvement locates at I=1. However, the consecutive frames also have some benefit since the overall disturbances also decrease. Therefore, we choose the minimum consecutive distances in the chosen N_{iter} frames to be 6 for G.723.1 and MPEG-4 CELP, which has an overall decrease in disturbance which is about half of that when I = 1. Results are shown in Table 5-6 and the FCB bits are reduced for G.723.1, but not in MPEG-4 CELP. One possible reason is that the inter-frame dependence is so small in MPEG-4 CELP so that preventing pulses from locating in close positions may increase the required FCB pulses to maintain the same quality.

Next, in Table 7-8 the results are performed by "max80decade+mindis" for G.723.1 and "max80decade" for MPEG-4 CELP, since they are the best results regarding both the bit-rate and execution time. In Table 7 we show the relative subjective measure (preference test) between the adapted speech and the standards. The relative subjective measure gives a subjective quality measure of an utterance U_1 relative to utterance U_2, which is defined as

$$G_{12} = \frac{1}{T} \sum_{t=1}^{T} g_{12}$$

$$g_{12} = \begin{cases} 100 & if\ U_1 \text{ is better than } U_2 \\ 0 & if\ U_1 \text{ is similar as } U_2 \\ -100 & if\ U_2 \text{ is better than } U_1 \end{cases} \quad (2)$$

where T is the total number of test. G_{12} ranges from –100 to 100, and it represents the percentage of people who prefer U_1 if $G_{12} > 0$ or dislike U_1 if $G_{12} < 0$. In this experiment 20 persons are involved and for each person eight 5-sec test sentences for each codec are given. From the table it is clear that the subjective speech quality of the adapted speech is slightly worse but almost the same with both the standards.

In Table 8 we list the bits reduction rate of FCB and overall codec in different types of speech. First we can see

that the reduction in male speech is more than that of female speech. It can be attributed to the fact that female speech contains more high-frequency un-predictable components, and thus more FCB pulses are needed to model the random excitation signal. Second, reduction rate in silences are much larger than in speech. In fact, the FCB pulses allocated in silences are almost equal to the minimum FCB pulses in both codecs. The reason is that by using the PESQ analysis, very few pulses are needed since the perceptual disturbances in silences are very small. Therefore the proposed method can largely reduce the required bit-rate in silences without the need of a VAD. Note that the reduction rate of silences in MPEG-4 CELP is smaller than that in G.723.1 because of the minimum FCB pulse is larger than G.723.1. On the other hand, the bit reduction rate in speech for MPEG-4 CELP coder is larger than that of the G.723.1 coder. The reason is that by adopting the perceptual disturbance analysis, fewer pulses are already sufficient in certain parts of speech. Therefore, MPEG-4 CELP coder with a maximum 10 pulses per subframe exhibits larger space to improve than G.723.1. Finally, the overall FCB bits reduction of the two codecs are 35.9% and 32.5%, which correspond to overall codec bits reduction 18.8% and 18.9%.

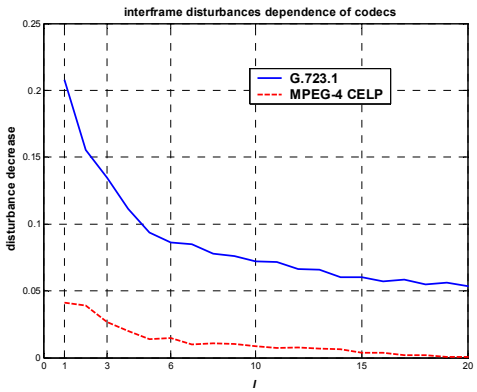

Figure 5: Statistic of the overall decrease of disturbance in I_{th} frame when we increase the pulse number at frame 1.

	FCB bits		Execution time	
	F	**M**	**F**	**M**
G.723.1	99	99	1	1
+ max 1	64	60	1411	1295
+max80decade+mindis	65	62	18	17

Table 5: Results of the case G.723.1 with distance limitation.

	FCB bits		Execution time	
	F	**M**	**F**	**M**
MPEG-4 CELP	120	120	1	1
+ max 1	87	75	1787	618
+max80decade	87	75	33	5
+max80decade+mindis	88	80	36	14

Table 6: Results of the case MPEG-4 CELP with distance limitation.

compared with	**Relative subjective measure**
G.723.1	-4.37%
MPEG-4 CELP	-1.25%

Table 7: Subjective measure of the proposed method.

		FCB speech	FCB silence	FCB overall	codec overall
G.723.1	**F**	21.5%	66.6%	34.3%	18.0%
	M	28.4%	65.8%	37.4%	19.6%
MPEG-4 CELP	**F**	24.2%	40.6%	27.5%	16.0%
	M	35.2%	40.8%	37.5%	21.8%

Table 8: FCB reduction rates of the proposed method.

5. Conclusion

In this paper, we have presented a pulse number allocation scheme for reducing the offline storage requirement of speech coders while maintaining the same speech quality. By exploiting the relationship between the disturbance value and PESQ score improvement and property of inter-frame dependence of speech codecs, the proposed scheme can achieve about 20% bit-rate reduction for both codecs. The proposed scheme is useful for reducing the offline storage requirement such as a corpus-based Text-To-Speech system or any voice data when they need to be stored on portable devices.

6. Acknowledgement

This paper is a partial result of Project A341XS1Q10 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan, R.O.C.

7. References

[1] Bishnu S. Atal and Joel R. Remde. A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates. *Proc. ICASSP*, pages 614-617, 1982.

[2] Schroeder, M. R. and Atal, B. S., "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," *Proc. ICASSP*, Mar 1985, pp. 937-940.

[3] ITU-T Recommendation G.723.1, "Speech coders: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s"

[4] Rao, A.V.; Ahmadi, S.; Linden, J.; Gersho, A.; Cuperman, V.; Heidari, R., "Pitch adaptive windows for improved excitation coding in low-rate CELP coders", *IEEE Transactions on Speech and Audio Processing*, Volume: 11, Issue: 6, Pages: 648 – 659, Nov. 2003

[5] Amada, T.; Miseki, K.; Akamine, M., "CELP speech coding based on an adaptive pulse position codebook", *Proc ICASSP*, March 1999, pp. 15-19

[6] Stachurski, J.; McCree, A.; Viswanathan, V.; Heikkinen, A.; Ramo, A.; Himanen, S.; Blocher, P., "Hybrid MELP/CELP coding at bit rates from 6.4 to 2.4 kb/s", *Proc ICASSP*, April 2003, pp 153-6

[7] Shlomot, E.; Cuperman, V.; Gersho, A., "Combined harmonic and waveform coding of speech at low bit rates", *Proc ICASSP*, 1998, pp. 585 - 588

[8] Paksoy, E.; Srinivasan, K.; Gersho, A., "Variable rate speech coding with phonetic segmentation", *Proc ICASSP*, 1993, pp. 155 - 158

[9] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs"

[10] http://www.tnt.uni-hannover.de/project/mpeg/audio/

[11] http://www.ldc.upenn.edu/Catalog/LDC93S10.html