CELP BASED SPEECH CODING WITH FINE GRANULARITY SCALABILITY

Fang-Chu Chen and I-Hsien Lee

Computer and Communications Research Laboratories
Industry of Technology Research Institute, Hsin-Chu 300, Taiwan, R.O.C.

ABSTRACT

General audio and video with *Fine Granularity scalability (FGS)* has become favored in the next generation multimedia coding standards due to its high flexibility in channel rate adaption. However, the FGS phenomenon has not yet been fitted into existed speech codecs. In this paper, we introduce the feature of FGS to the *Code Excited Linear Prediction (CELP)* based speech coding algorithm by adjusting the amount of transmitted fixed excitation information. Besides, we further improve the algorithm by relaxing the constraints and re-ordering the pulses sequence. To achieve this target, we need to make modifications on the conventional coding algorithm but the computation overhead and affected module is little. As a consequence, the developers can easily migrate their on going codec to one with the FGS advantage in a short time.

1. INTRODUCTION

Ever since the multimedia data was highly used in streaming transmission, the coded data quality is no longer the only criterion in multimedia signal compression. To ensure that all the information can be received in the decoder side, the amount of coded data should be no more than the available bandwidth. However, the bandwidth usage in a transmission channel can not always be observed at the encoder side. As a consequence, improperly compressed data could be lost if the traffic is congested. This problem can be solved by layer coding, i.e., a scalable bit stream consisting of a *base* layer followed by one or several *enhancement* layers. The base layer of which is the minimum requirement and has to be received by the decoder in order to maintain an acceptable quality of the decoded contain of the stream. The enhancement layers, on the other hand, are used to improve the media quality and it can be ignored one layer at a time.

In addition to layer coding, *Fine Granularity Scalability (FGS)* [1] is a new approach, allowing the bitstream to be discarded with finer granularity instead of a whole layer. FGS provides the channel traffic supervisor a much easier and more flexible way to control the traffic. General audio and video coding algorithms with FGS have been adopted as part of MPEG-4 international standard [2]. However, an FGS speech coding technique has not yet been standardized. The FGS algorithms used in MPEG-4 general audio and video share a common strategy, that the enhancement layers are distinguished by the different bit significance level at which a bit plane or a bit array is sliced from the spectral residual. This method, however, does not work well for a highly parametric speech coder such as CELP-based ITU-T G.723.1, and AMR in 3GPP [3][4].

Hence, we propose a new scheme in the attempt to provide the CELP based speech codec with the FGS feature. The proposed scheme is developed by the observations mentioned in [5] and [6]. In [5], the author proposes that the excitation of certain subframe can be generated just by the extension from the previous subframe with little performance scarification. That means the stochastic pulses of these subframe can be ignored in speech synthesis. This concept is to alternate the methods between CELP and *Self-Excitation Linear Prediction (SELP)*[7]. The authors of [6] also report that these pulses can be added back , one at a time, to the subframes by following proper modification. Adding back these pulses can increase the re-constructed speech quality and this also implies that the granularity of the bitstream is in single pulse basis. Furthermore, this feature can also be the fundamental technique of voice band embedded data application [8]. However, there exists some constraints on the original FGS approach proposed by [6] and these will be discussed and solved in section 3. Before that, we would like to introduce the details of the original approach first.

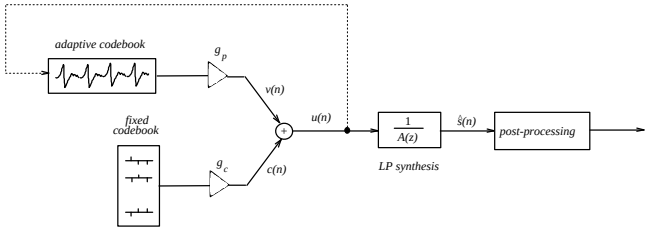

Figure 1: The combined excitation vector and its feedback path in CELP algorithm [4].

2. CELP BASED SPEECH CODING AND ITS FGS APPROACH

In a practical CELP-based speech coder, *Linear Predictive Coding (LPC)* model with only *adaptive codebook* always leaves errors between the synthesized speech and the original one. In a common process, the errors due to the imperfections of the model are compensated by stochastic, *fixed codebook*, process. The stochastic process is often time implemented by fixed-code pulses which are added to the pitch part of the excitation. Then, the combined excitation vector is filtered through the LPC filter and the errors can therefore be minimized. Specifically, speech component generated by the fixed-code pulses is used to enhance the quality of synthesized speech in subframe basis as can be seen in Fig. 1.

Parameters coded	sub-frame 0	sub-frame1	sub-frame 2	sub-frame 3	Total
LPC indices					24
Adaptive code-book lags	7	2	7	2	18
All the gains combined	12	12	12	12	48
		8		8	40
Pulse positions	12	12	12	12	48
		0		0	24
Pulse signs	4	4	4	4	16
		0		0	8
Grid index	1	1	1	1	4
		0		0	2
TOTAL:					158
					116

Figure 2: Bit allocation of low-rate coder of G.723.1 and its reduction in [5].

2.1. Bit Rate Reduction Method

In some cases, the channel bandwidth could be too small and this would limit the transmission of coded bitstream. Properly adapting the coding algorithm might be the solution but it is not an efficient way. According to CHEN[5], the number of the fixed-code pulses, which occupies a big percentage of the total bit rate, can be cut in half by ignoring those pulses in the odd-numbered subframes. As can be seen in Fig. 2, the method leads to a 27% bit-rate reduction in low-rate coder of ITU-T G.723.1 with only 1 dB *Segmental Signal-to-Noise Ratio (SEGSNR)* degradation in the decoded speech. As a consequence, the limited bandwidth might be satisfied by following this modification.

Since this procedure drops the pulses information in odd numbered subframes, the pulses searching procedure of these subframes is skipped in this bit-rate reduction approach. In other words, pulses information in odd-numbered subframe will not be received or take part in speech synthesis. The total excitation of odd numbered subframe is constructed just by the excitation of the previous subframe. Then, this excitation is feedback to the adaptive codebook, by following the dash line in Fig. 1, for the incoming speech synthesis. All the other parameters and their related coding procedure is not affected by this modification, hence, this technique only requirs a minor modification if the coder is already being established.

Base on the previous study [6], FGS can be achieved by searching the pulses of the odd-numbered subframes and delicately adding them back to the bitstream. In other words, the information bits associated with the fixed-code pulses of the odd-numbered subframes can be viewed as the enhancement layer of the stream. The following subsection describes the details of the modifications involved in realizing this concept.

2.2. CELP Based FGS

The enhancement layer of an FGS bit stream is allowed to be discarded as a whole or by part depending on the transmission environment. Placing the odd-numbered subframe pulses in the enhancement layer implies that the number of those pulses received by the decoder is unknown at the encoder side. The purpose of the analysis-by-synthesis method, by imbedding a decoder in the encoding process, is for the encoder to foresee the exact speech decoded by the decoder on the other end of the transmission line. If the encoder has no knowledge about the number of odd-numbered subframe pulses actually used by the decoder, it would have no base for constructing the best parameters to be sent to the decoder. This phenomenon could jeopardize the analysis-by-synthesis method used in the standard coder.

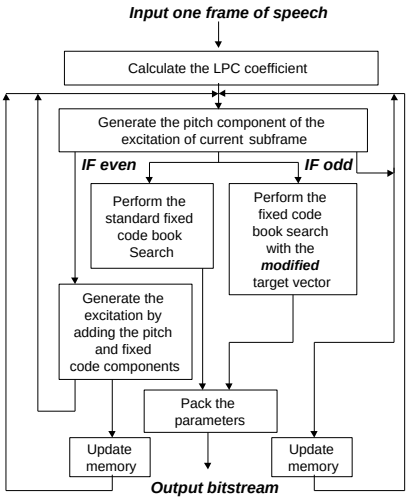

Figure 3: The flow chart of the modified encoder shown in [6]

One way to minimize this problem is to assume the worst case of the receiving condition, i.e., always assume that the decoder receives none of the information bits from the enhancement layer. To be more precise in terms of implementation, the excitation vector and the memory states (of the LPC filtering) passed over from an odd-numbered subframe to the next even-numbered subframe have to be constructed without any information from the odd-numbered subframe pulses as can be seen in Fig.3. The odd-numbered subframe pulses are still searched and generated, however, they are purely used for extra quality enhancement of that subframe and are never recycled in the future subframes. If the encoder is allowed to recycle any of the odd-numbered subframe pulses which are not received by the decoder, then, the coded parameters selected for the next subframe might not be the right choice for the decoder and an error would occur.

3. GENERALIZATION OF CELP BASED FGS

In the aforementioned discussion, the pulses information of the coded bitstream can be removed, one at a time, to achieve the phenomenon of FGS. However, this technique still has some problems that might restrict the usage of itself. As we can see in Fig.3, the pulse information of enhancement layer is confined to the odd-numbered subframes. Hence, the amount of data allowed to be dropped is limited, which limits the ability in channel bandwidth adaption. In this section, we introduce the schemes that relax the constraints and provide the coded bitstream with a higher performance in either bandwidth adaption and re-constructed speech quality.

3.1. Relaxing the Constraints of CELP Based FGS Coding

In the coder of *Wide-Band Adaptive Multi-Rate (WB-AMR)*[4], most the coding modes share the same coding procedure except the number of pulses. Amoung these modes, each subframe contains 2 to 24 pulses which determines the bit rate of the coded

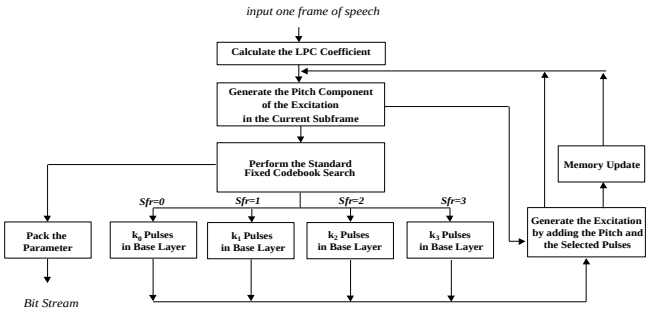

Figure 4: The flow chart of the generalized FGS encoder.

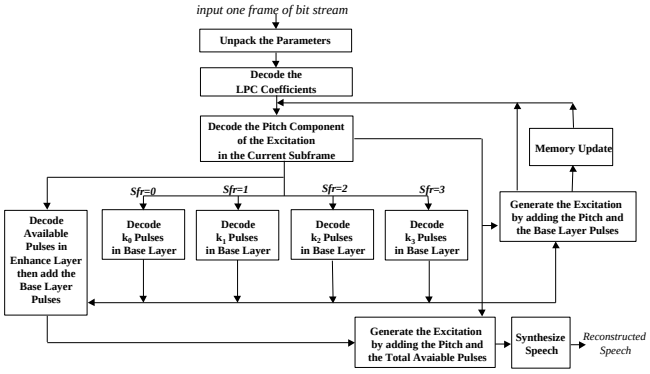

Figure 5: The flow chart of the generalized FGS decoder.

streams. FGS can be achieved by simply removing the pulses in odd-numbered subframe and keeping all the other coding procedure in the original way. However, this approach does not allow a big enough range for bit rate scalability due to the constraints of enhancement layer arrangement. In order to ease this problem, we make some modifications on this conventional procedure by the features listed as below:

1. The pulses in the enhancement layer are not confined to odd-numbered subframe.
2. The number of pulses in the enhancement layer can be smaller or equal to the total number of pulses in that subframe.

Generally speaking, each frame of the speech signal to be encoded is partitioned into 4 subframes. Let's assume that, all the subframes contain N pulses as their stochastic excitations. In the proposed approach, the pulses belonging to the enhancement layer can be distributed in all the subframes and subframe *i* contains k_i pulses for its enhancement layer. This value should be predetermined before the speech analysis because we need to make a worst case assumption to ensure the analysis-by-synthesis method. Hence, the number of pulses in the base layer from subframe *i* is $N - k_i$ and there are only $N - k_i$ pulses added to the adaptive excitation. That means only these pulses would take part in the excitation buffer refreshing and synthesis filter memory updating procedure as can be seen in Fig.4 and Fig.5.

Although the bit rate control is much similar to the approach mentioned in section 2, this work is nothing to do with a subframe basis CELP/SELP switching coding. Since all the subframes may contain some pulses to be their base layer information, the synthesized speech can be continuously enhanced by these pulses. This ensures that the speech quality will not be dominated by the SELP based synthesized interval due to all the subframe having additional stochastic pulses in its base layer. As a consequence, we can achieve a higher re-constructed speech quality by following the proposed method.

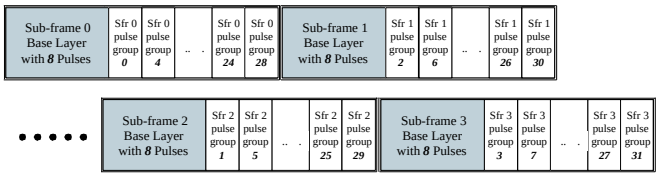

Figure 6: Pulses removing sequence in the proposed FGS scheme.

Furthermore, re-constructed speech quality is not only affected by the number of received pulses in the enhancement layer but the predetermined value of $K_{total} = \sum_{i=0}^{3} k_i$. A larger value of K_{total} represents a larger amount of data is arranged as the enhancement layer. This implies that a higher ability in bandwidth adaption because more pulses, K_{total}, can be dropped under this assumption. However, this also makes the error minimization procedure into a worse case because of the smaller amount of pulses that can be used in memory updating. The determination of K_{total} becomes a trade-off choice in the proposed method and this can be decided depends on the minimum quality requirement or channel bandwidth variation. Hence, the proposed method offers a higher flexibility in speech coding.

3.2. Re-ordering the Pulses Removing Sequence

A full-length bitstream contains a base layer and a full-length enhancement layer. The criterion is to transmit those bits needed in the base layer before those for the enhancement layer. Furthermore, the bit order in the enhancement layer has to be modified in order to accommodate the ability of flexible bit rate transmission. In [6], the bits for the pulses of one odd-numbered subframe are grouped together. With this ordering, bandwidth supervisor can discard the pulses in the same odd-numbered subframe first before affect the pulses in the other subframe.

According to the previous discussion, a full-length enhancement layer contains K_{total} pulses which might locate in all the subframes instead of only in the odd-numbered ones. Although the pulses in these subframe can also be removed by following the method mentioned above, we will further propose a new way that can improve the re-constructed speech quality in the adaptive bit rate range. Since the sequence of pulses dropping is also an important implementation criterion in FGS speech coder, the proposed idea is that the pulses in different subframes can also be removed at the same time.

As can be seen in Fig.6, each subframe contains 16 pulses and only 8 pulses are arranged as the base layer information. If the bandwidth capability only allows the base layer and 16 enhancement layer pulses to pass through the channel, 16 pulses in the enhancement layer should be removed and these pulses are equally distributed in all the subframes. Specifically, all of the subframes ignore the last 4 pulses to achieve this target. The indexes of pulses represent the sequence of pulse dropping if the bandwidth is not enough. A larger value means this pulse will be dropped first. In our study, dropping the pulses by following the sequence men-

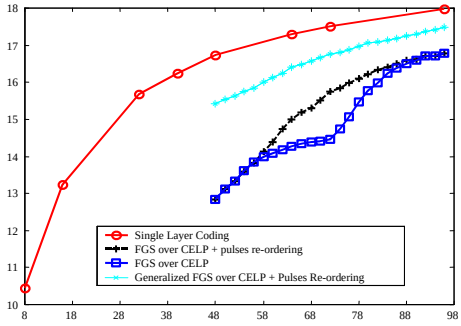

Figure 7: SEGSNR of the re-constructed speech in different approaches.

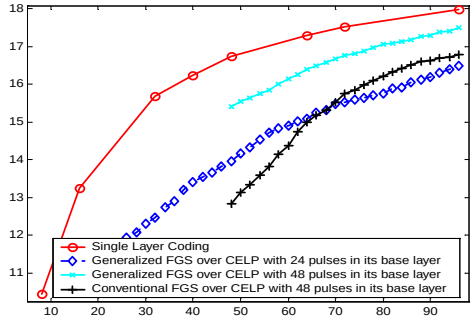

Figure 8: SEGSNR of the re-constructed speech with different base layer length.

tioned in Fig.6 is a more proper approach in maintaining the re-constructed speech quality and the simulation result is shown in the following section.

4. SIMULATION RESULT AND ITS APPLICATIONS

In our design, we take the specification of WB-AMR [4] as the standard coder and the proposed FGS approach is developed based on this specification. Although an FSG speech coding technique only involves very little modifications from the standard methods, the amount of base/enhancement layer data should be pre-defined before the speech coding procedure. As have been mentioned in section 3, the highest bit rate mode of the standard coder contains 24 pulses in each subframe. In the following demonstration, half of the pulses are selected as the enhancement layer information in order to compare the re-constructed speech quality with the original FGS approach.

Fig.7 shows the curves of re-constructed speech quality in various approaches. The x-axis represents the total number of pulses in a single frame and the y-axis is the *Segment Signa-to-Noise Ratio (SEGSNR)*. All markers upon the curve represent the points that can be selected as the available bit rates. As we can see in the figure, although the FGS approaches have a poor performance in re-constructed speech quality, these approaches can offer a higher ability in bit rate adaption. Furthermore, there is no need to adapt the coding mode if the channel bandwidth is changed in the FGS design.

Because the FGS approach in [6] discards all the pulses on an odd-numbered subframe case. Speech quality of this FGS approach, square marked curve, suffers a saturated like result while only half of the enhancement layer pulses are received. This problem can be solved by using the enhancement layer pulses re-ordering scheme. That means we can achieve a higher quality if we equally remove the amount of enhancement layer pulses in both of the odd-numbered subframes. The curve with cross marker demonstrates that we can improve $1.5dB$ speech quality by simply dropping 12 pulses in these subframes instead of 24 pulses in a single odd-numbered subframe.

The quality can be further improved by following the proposed FGS approach. In Fig.7, the curve with double cross markers is generated by using the proposed FGS approach and the pulses re-ordering scheme. Because all the subframes contain 12 pulses as their base layer information, the analysis-by-synthesis procedure can be processed in a smoother way. That means the re-constructed speech quality will not be dominated by the subframe which is predicted by SELP algorithm. In our simulation, SEGSNR value of this approach is only $1dB$ lower than the single layer coding. This approach also has a $2.5dB$ advantage compared with the origin CELP based FGS scheme. Although simply configuring the length of base layer will scarify the re-constructed speech quality, this approach may increase the ability in bandwidth adaption as can be seen in Fig.8. As a consequence, the parameter should be decided to satisfy the required rate adaption range and the lowest demanded speech quality.

5. CONCLUSION

In this paper, we introduce one way in achieving CELP based FGS speech coding. Furthermore, we also propose two ways in generalizing it to have a higher ability in bandwidth adaption and better re-constructed speech quality. Amount these methods, only little modification is required and the computation has almost the same load in the proposed algorithm. As a consequence, we may migrate the original codec to achieve the FGS phenomenon in a short time.

References

[1] W. P. Li, "Overview of Fine Granularity Scalability in MPEG-4 Video Standard," *IEEE Trans. On Circuits and Systems for Video Technology*, Vol. 11, No. 3 March 2001.

[2] ISO/IEC 14496, the MPEG-4 standard.

[3] ITU-T Recommendation G.723.1: 'Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s.'

[4] 3GPP TS 26.190, 'Speech Codec speech processing functions; AMR Wideband speech codec; Transcoding functions (Release 5).'

[5] F. C. Chen ,"Suggested new bit rates for ITU-T G.723.1", *Electronics Letters*- Vol. 35 No.18 p. 1523, 1999.

[6] F. C. Chen, I. H. Lee, "Speech Coding with Fine Granularity Scalability Based on ITU-T G.723.1", *International Computer Symposium (ICS)*, Hualien, Dec. 2002.

[7] J. I. Lee, C. k. Un,"Multistage Self-Excited Linear Predictive Speech Coder", *Electronics Letters*- Vol.25 p.1249-1251, 1989.

[8] F. C. Chen, I. H. Lee,"Introduction of AMR-WB Submodes Based on Fine Granularity Scalability", *3GPP TSG-SA4#23 Meeting*, Tdoc S4-020501, Montreal, Oct. 2002.