EX-CELP : A SPEECH CODING PARADIGM

Yang Gao, Adil Benyassine, Jes Thyssen, Huan-yu Su, and Eyal Shlomot
Conexant Systems, Inc.
Email: [yang.gao, adil.benyassine, jes.thyssen, huan-yu.su, eyal.shlomot]@conexant.com

ABSTRACT

This paper presents the core technology of novel enhancements to traditional CELP coding, coined eXtended CELP (eX-CELP). It is centered on a combined and selective usage of closed-loop/open-loop approach, and variant algorithm structure concept. The above two concepts are complemented by new features and refined existing technologies. The eX-CELP paradigm was used in several speech coding systems. It is the core technology of the recently chosen candidate for the 3G-CDMA speech codec standard. It was the best candidate for ITU-T 4kbps codec qualification test, and became the basis technology for a consortium candidate to the ITU-T 4kbps speech coding competition.

1. INTRODUCTION

CELP coding of speech signals has been proven to deliver toll quality or near toll quality at high to medium bit rates. Yet, until recently, it is still challenging to achieve such quality at a bit rate as low as 4 kbps. This paper describes a successful attempt to realize this achievement based on an extended paradigm of CELP. Although the eX-CELP system was originally developed to achieve toll quality at 4 kbps, our experiments and test results showed that this technology is also successful and suitable for both high and medium bit rates. Fig.1 and Fig.2 illustrate the basic structure of the eX-CELP encoder and decoder.

One of the main themes of the eX-CELP technology is the judicious combination of the closed-loop approach and the open-loop approach, together with a careful selective usage of them. This mechanism is coined COLA, and its main objective is to intelligently employ the most appropriate approach for different types of input signals in order to preserve the perceptually important contents. Another important feature in the eX-CELP technology, termed Variant Algorithm Structures (VAS), is introduced to handle different kinds of speech signal more efficiently using a safe "soft" decision mechanism for the selection of the appropriate algorithm structure. Here, soft decision means to completely or partially use closed-loop or delayed information to make reliable decisions. The two concepts COLA and VAS are used throughout the eX-CELP system even for high bit rate coders. Reflecting upon the above mentioned concepts, several techniques will be proposed in Section 2, Section 3, and Section 4.

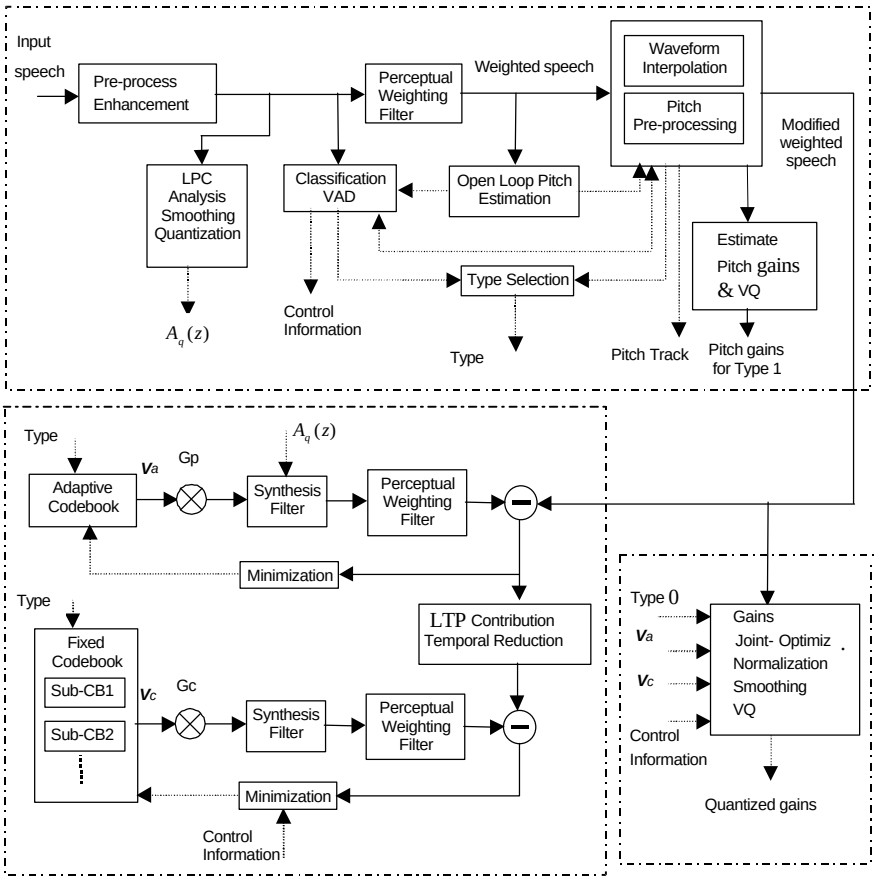

Fig. 1 eX-CELP encoder

Section 2 describes pitch related techniques. Similar to the RCELP concept [1], a new pitch-preprocessing algorithm is carried out that modifies the weighted speech (see Fig.1) rather than LPC residual for matching a coded pitch track in a fast computational way. Pitch gain in voiced transition areas is enhanced through harmonic smoothing (see Fig.1) or waveform interpolation. After applying these open-loop techniques, the pitch processing is further refined for irregular signals, using the closed-loop concept.

Section 3 is dedicated to the fixed codebook structure and the associated searching techniques. The fixed codebook is such that it is made up of several sub-codebooks (see Fig.1), benefiting from the principle of VAS. First, the best sub-codebook is selected in a COLA fashion. Then, this sub-codebook is furthermore searched in a fast iterative closed-loop way. The second target signal for fixed codebook searching is perceptually modified (see Fig.1) in an open-loop procedure to improve the search performance.

Section 4 addresses the gain related techniques that are also in the sprit of the concepts of COLA and VAS. Tailored gain quantization schemes (see Fig.1 and Fig.2) are used for different types of speech signals. For

stationary voiced speech, pitch gains are pre-estimated and pre-vector-quantized in an open-loop fashion prior to the subframe processing. However, fixed codebook gains are determined and delay-vector-quantized in a closed-loop fashion. For the rest of the signals, all gains are calculated at first as in a traditional closed-loop approach, then they are further normalized, smoothed and quantized in a COLA fashion.

All of the above mentioned features are in line with the concepts of COLA and VAS, hence the eX-CELP system boasts two basic types of algorithms and bit allocations, referred as Type 0 and Type 1 (see Fig.1 and Fig.2).

1. ENHANCING PITCH CONTRIBUTION

Maximizing pitch prediction gain, whenever appropriate, at a low cost is crucial to low bit rate coders. Pitch pre-processing is used to reduce the bit rate to code pitch lag information. It could be performed on two "extreme" conditions: 1) in LPC residual domain in an open-loop fashion or 2) in perceptually weighted signal domain in a closed-loop fashion. The former approach is simpler but the quality might not be optimum. The latter approach could provide the best quality at the expense of a huge complexity increase. The eX-CELP system introduces a new pitch-preprocessing algorithm that produces a quality very close to the second approach but enjoys the low complexity of the first approach.

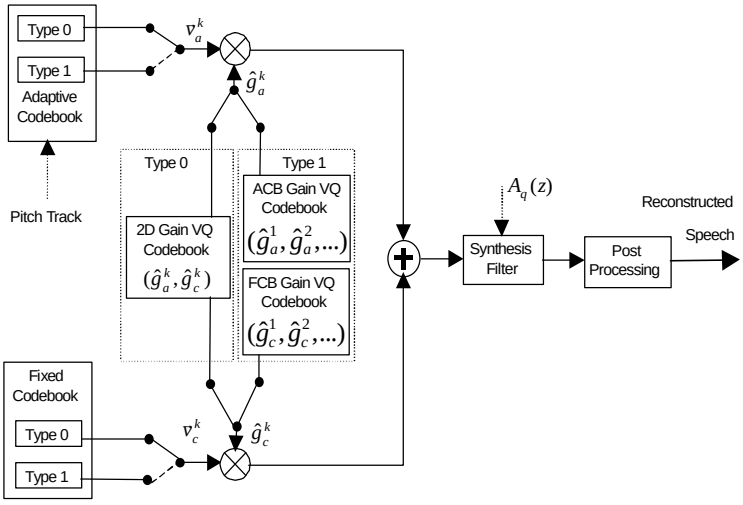

Fig. 2 eX-CELP decoder

The proposed pitch preprocessing algorithm modifies perceptually weighted speech rather than LPC residual signal. Specifically, a continuous warping is always applied to voiced area in order to avoid any possible discontinuity. According to typical closed-loop principle, one can imagine that all the warping candidates must be evaluated and tested in order to find the best one; however, this leads to heavy computational requirements. To dramatically reduce the complexity, a cost window function is proposed to multiply with the target signal during searching for the best local delay, as shown in the Fig. 3. The maximum value of the cost function is 1.0 and the smallest value must be smaller than 1.0 (even set to 0.0). During a first step, an integer delay is searched by maximizing the correlation between the current segment signal and the windowed target signal. Then, the correlation values are up-sampled to find the precise local delay D. Finally, the low energy area between two pitch peaks is warped with a linearly variable local delay from 0 to D, whereas, the energy peak area is shifted by delay D.

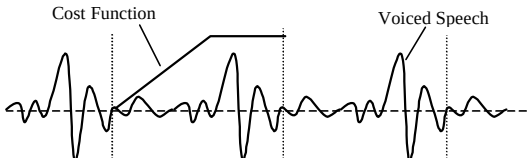

Fig. 3 Multiplying a cost function to the target signal during searching for the best local delay.

At low bit rate, increasing the pitch contribution in voiced transition areas is always beneficial. Unfortunately, the amount of modification required might be substantial, making it difficult to achieve by employing the pitch-preprocessing without running the risk of degrading the quality of the original speech. Therefore, it is not suggested in the eX-CELP system to only rely on pitch-preprocessing to handle voiced transition areas. In fact, voiced transition areas exhibit not only rapid pitch lag changes, but also irregular waveform shapes. Neither pitch preprocessing nor traditional LTP approach could provide high LTP gains for these areas at low bit rate. In order to achieve this goal at low bit rate, a harmonic smoothing or waveform interpolation is proposed to pre-smooth voiced transition areas in an open-loop fashion. This will artificially enhance the pitch correlation. As a consequence, this will result in a faster building-up of the adaptive codebook for onsets.

After finishing the harmonic smoothing for some voiced transition areas and pitch preprocessing for all the voiced speech areas, a delayed decision will be made to choose Type 0 or Type 1 coding scheme. Subsequently, different gain systems are used respectively for the two types. Moreover, for Type 0, traditional closed-loop pitch prediction is added to further improve pitch prediction.

3. MULTI SUB-CODEBOOK STRUCTURE AND FAST SEARCH

Another typical aspect of the VAS principle with a safe soft decision is the use of a multi-sub-codebook structure for the fixed codebook. At high bit rate, a large single structure fixed codebook could probably fit all kinds of possible signals. However, at low bit rate, it is very difficult to keep the naturalness and robustness of the perceived quality with only one structure of the fixed codebook. The concept of multi-sub-codebooks is different from that of multi-mode excitation where the best mode of excitation is selected using a "hard" decision of classification. With multi structure codebook, the best sub-codebook is selected based on the COLA principle with a safe "soft" decision. What this means is that the closed-loop waveform matching criterion is used to select the best sub codebook with the assistance of the open-loop classification information. Once the best sub-codebook is chosen, some fine searching for the best code-vector is done within this specific sub-codebook. Our experiments showed that

high bit rate coder also benefits from the multi-structure codebook.

In the eX-CELP system, there are usually three typical sub-codebooks. The first sub-codebook is aimed to represent relatively more periodic voiced speech signal. This sub-codebook contains a number of pulses with good position (or temporal) resolution. Usually the number of pulses in this sub codebook is less than other sub codebooks. In general, each possible sample position in the subframe must be covered at least once by one pulse position candidate. For this sub-codebook, not only forward pitch enhancement is applied; but also a novel backward pitch enhancement is introduced as shown by the impulsive response in Fig. 4.

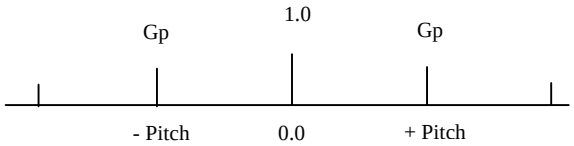

Fig. 4 Response of Forward and Backward Pitch Enhancement

The second sub-codebook still provides pulse-like code-vectors, which tries to code energy-pulse-like signals that the first sub-codebook could not match well. In our design, a Dynamic Track Structure (DTS) is explored in this sub-codebook. DTS refers to the fact that position tracks for some of the pulses or all the pules are assigned dynamically relative to reference positions which are available in the decoder, so that fewer bits are needed to code these pulses. The reference positions could come from the few pulse positions coded with fixed tracks (not DTS) or from analyzing the adaptive codebook energy distribution available in the past excitation, for example. The pitch enhancement is still applied for this sub-codebook; but often using a weaker enhancement gain.

The third sub-codebook is designed to handle signals that either exhibit stable energy distribution of the LPC residual within a subframe or are of the noise-like nature. This sub-codebook contains code-vectors that have almost uniformly distributed energy within a subframe. For a high bit rate coder, the sub-codebook could be simply constructed by putting more pulses and limiting each pulse position to a small range, which prevent all the pulses from crowding in a small area. For a low bit rate coder, it is a good solution for this sub-codebook to include Gaussian noise vectors. A pulse-like sub-codebook is usually more favored by the closed-loop criterion than a noise-like sub-codebook, therefore, when the open-loop classification shows that a noise-like excitation is needed, it is better to favor the noise-like sub-codebook to be selected by modifying the closed-loop criterion.

Besides the merits of the structure of the fixed codebook, the search performance of the codebook is also important. It typically depends on two aspects. One is to have a perceptually good target signal. A well-known approach to get that is to employ a perceptual weighting filter. Another aspect is related to fast searching of the fixed codebook. There already exist many fast search approaches. However, they may not be suitable for multi-structure codebook.

Regarding the first aspect, a simple way to perceptually improve the target signal for fixed codebook search is to temporally de-emphasize the LTP contribution:

$$T_{FCB}(n) = T_{total}(n) - l\,(R_P)\cdot G_P \cdot F_{AC}(n) \qquad (1)$$

where $T_{FCB}(n)$ is the target for fixed codebook search, $T_{total}(n)$ is the total target, G_P is the pitch gain, $F_{AC}(n)$ is the filtered adaptive codebook signal, and $l\,(R_P)$ is the de-emphasizing factor which is a function of the normalized correlation between $T_{total}(n)$ and $F_{AC}(n)$: $R_P = \langle T_{FCB}(n),\ F_{AC}(n) \rangle$. If the long term prediction results in a perfect waveform matching (R_P is close to 1.0), the de-emphasizing factor is set close to 1.0; otherwise the de-emphasizing factor is set to a smaller value. The final fixed codebook and adaptive codebook gains will be re-determined after finishing the fixed codebook search. The reason why the perceived quality can be improved by doing this can be summarized by two points. The first point is that the two excitation components (adaptive codebook and fixed codebook) are not jointly optimized due to the complexity and the de-emphasizing factor can somehow balance the contribution from the two components. The second point is that the de-emphasizing factor actually makes the second target signal reserve more information in the peak energy area and become more periodic in the voiced area.

Regarding the second aspect, in order to make the multi-structure codebook search faster and more efficient, an iterative way to search for the pulse codebook is proposed; one iteration is usually performed by moving one pulse position and a group of iterations with that every pulse is searched for once is referred as a “turn” which provides a candidate codeword from the fixed codebook. Basically the current search turn tries to improve the result obtained from the previous turn. The larger the number of turns, the better quality obtained. All the search turns are based on originally defined tracks or a subset of the original tracks. The subset must be determined for each subframe before the iterative search is started. With the above approach, the best sub-codebook can be selected first according to the result obtained from the first turn. Then, further turns are performed on the selected sub-codebook to improve the search performance.

The algorithm for each turn is designed in terms of the complexity requirement. One simple solution for each turn is to sequentially search for each pulse position (and/or its sign) one pulse after another from the first pulse to the last pulse by temporally fixing the influence from all other existing pulses, excluding the current pulse. For the first turn, the influence comes from the previous positioned pulses and the influence from the following pulses is temporally set to zero. During the second turn, the search for any current pulse has a temporally fixed influence from all the other pulses. Successive turns will repeat in the same way. The quality will be improved from one turn to the next as the influence condition from other positioned pulses to the current pulse is changed. This can be highlighted more by examining the search criterion:

$$Maximize \quad \left\{ \ (B \cdot C^{T})^2 \,/ (C \cdot \Phi \cdot C^{T}) \ \right\} \qquad (2)$$

where $\boldsymbol{B}$ is transformed target, $\boldsymbol{F}$ is correlation matrix and $\boldsymbol{C}$ is a code-vector, the numerator and denominator of the criterion in the searching loop can be very simply computed, modified, or updated in an iterative procedure as only one index for one pulse

changes in the loop. To save possible large memory due to $\boldsymbol{F}$, we can even compute the denominator without using $\boldsymbol{F}$ in a fast iterative way. A little more complex solution could be that for each searching turn, all the pulses are grouped into several pairs of pulses and each pair of pulses are jointly optimized from one pair to another.

4. NEW STRATEGY OF GAIN DETERMINATION

This section describes efficient approaches for determining adaptive codebook and fixed codebook gains in the eX-CELP system. Here again, we benefit from the concepts of COLA and VAS. At high bit rate, the traditional way to evaluate and quantize gains in a closed-loop approach does not show significant problems. However, at low bit rate we can see the deficiency of this method that not only takes a larger percentage of total bits to quantize gains, but also generates non-stable synthesized speech signals. The following design is to overcome this deficiency.

First, a pre-determination of the pitch gains for stationary voiced speech is employed. Stationary voiced speech corresponds to Type 1 in the eX-CELP system. This approach is motivated by the fact that the closed-loop pitch gains in a voiced speech area could become unstable due to waveform mismatch mainly when the subframe is located between two pitch peaks. Since the open-loop pitch gains evaluated from original weighted speech signal end up to be more stable in stationary voiced speech area, we can easily replace the closed-loop pitch gains with the open-loop pitch gains and benefit from the stability to do pre-VQ of them using much fewer bits. The determination and quantization of the pitch gains is performed before the subframe processing. Subjective tests reveal that only 4 bits to code 3 pitch gains for 3 subframes are needed to get transparent quality. Furthermore, 4 pitch gains can be coded with as low as 5 bits. The reason that the pre-determination approach of pitch gain works well could be because of the following reason: (1) the low energy area between two pitch peaks is relatively and perceptually less important, so that the perceptual influence of the mismatching between the closed-loop and open-loop gains is small; (2) forcing the use of a stable pitch gain could enhance the periodicity of synthesized speech; (3) a second excitation component from the fixed codebook could compensate for small deficiencies of the LTP contribution; (4) as the pitch gain is pre-quantized, the current pitch gain instead of one from previous subframe is used for pitch enhancement of the fixed codebook excitation. The fixed codebook gains are jointly quantized in a delay-decision using predictive VQ. For example, transparent quality for Type 1 speech is achieved by using 8 / 10 bits to code 3 / 4 fixed codebook gains.

For a Type 0 speech frame, the above method could also be adopted with more bits than Type 1, but still less than that would be required with the traditional way. The eX-CELP system rather employs the traditional joint quantization of the gains as in ITU-G.729. The difference here is that the closed loop gains are further modified and smoothed in an open-loop way through an energy normalization procedure. The smoothing degree may follow the stability of original signal. The smoothing could even operate respectively to adaptive codebook gain and fixed codebook gain.

5. SOME RESULTS FROM FORMAL TESTS FROM INTERNATIONAL COMPETITIONS

Conexant has participated in a number of recent international standard development processes. All our candidates are based on the eX-CELP core technology although the bit-rates, bit allocations and even the systems could be different.

One of the most recent results are from the 3GPP2/cdma2000 SMV (Selectable Mode Vocoder) standardization. The SMV coder is a variable bit-rate coder that contains three modes, which represent three different average bit rates (ABR). In the following table, the ABR of the existing standard EVRC is noted as E (for conversational speech, E is roughly around 4 kbps). Other ABR is defined as a fraction of E. Here is the average MOS achieved by the eX-CELP/SMV coder for each mode:

Coder	EVRC	Mode 0	Mode 1	Mode 2
ABR	E	1.0E	0.71E	0.56/0.6E
MOS(Clean)	3.581	3.900	3.636	3.464
MOS(Noisy)	3.346	3.569	3.528	3.526

For Mode 2, two average bit rates are defined, 0.56E for clean speech, and 0.6E for noisy speech. The eX-CELP/SMV coder won the highest average MOS for each mode respectively with clean and noisy speech conditions and become a chosen candidate [2].

During the last ITU-T coordinated qualification test for the 4kbps standard development, many candidates were submitted. The eX-CELP/ITU-T 4kbps coder was the only one that passed more than half of the requirements. Using the same eX-CELP/ITU-T 4kbps coder as the base-line, we have submitted a selection phase candidate generated by the consortium formed with AT&T, Conexant, Deutsche Telekom, France Telecom, Matsushita, and NTT [3], it is one of the only two retained candidates that moved forward for a fixed-point selection.

6. CONCLUSION

This paper presented the eX-CELP paradigm. The eX-CELP system is conceptually based on two main principles called COLA and VAS. These driving principles are behind the refinements and novelties as applied to pitch contribution, fixed codebook structure, gain determination and quantization, and two types of coding algorithms. The eX-CELP system achieved the best average quality of speech source coding during recent international competitions.

7. ACKNOWLEDGEMENT

The authors would like to thank Carlo Murgia for his effort to optimize the programming code.

8. REFERENCES

[1] W. B. Kleijn, R. P. Ramachandran, and P. Kroon, "*Generalized Analysis-by-Synthesis Coding and its Application to Pitch Prediction*", in Proc. Int. Conf. Acoust., Speech, Signal Processing, 1992, pp. I337-I340.

[2] Y. Gao et al., "*The SMV Algorithm Selected by TIA and 3GPP2 for CDMA Applications*", to appear in ICASSP2001.

[3] J. Thyssen et al., "*A Candidate for the ITU-T 4 Kbit/s Speech Coding Standard"*, to appear in ICASSP 2001.