THE SMV ALGORITHM SELECTED BY TIA AND 3GPP2 FOR CDMA APPLICATIONS

Yang Gao, Eyal Shlomot, Adil Benyassine, Jes Thyssen, Huan-yu Su, and Carlo Murgia

Conexant Systems, Inc.
Email: [yang.gao, eyal.shlomot, adil.benyassine, jes.thyssen, huan-yu.su, carlo.murgia]@conexant.com

ABSTARCT

During the years 1999 and 2000, the Telecommunication Industry Association (TIA) and the 3rd Generation Partnership Project 2 (3GPP2), managed a competition and a selection process for a new speech coding standard for CDMA applications. The new speech coding standard, which is coined Selectable Mode Vocoder (SMV), will become a service option in CDMA systems such as IS-95 and cdma2000, providing a higher quality, flexibility, and capacity over the existing speech coding service options, IS-96C, IS-127, and IS-733. Eight companies submitted candidates to selection phase. For all of the 36 test conditions, the Conexant SMV candidate was ranked at the top, or was statistically equivalent to other top-ranking candidates. The Conexant SMV candidate was chosen as the core speech coding technology for the SMV system. This paper describes the SMV algorithm developed by Conexant.

1. INTRODUCTION

The current CDMA communication standard, IS-95, and the future standard, cdma2000, include several variable-rate speech coding service options. Variable-rate speech coding algorithms are particularly important in CDMA systems, which use power control to achieve increased capacity as the average bit rate (ABR) is reduced. The first service option, IS-96C, operates at the Rate Set 1. Rate Set 1 consists of the rates 8.55 kbps, 4.0 kbps, 2.0 kbps, and 0.8 kbps (called full-rate, half-rate, quarter-rate, and eighth-rate, respectively). The quality of IS-96C was not competitive with other wireless standards or with wireline telephone services, and was replaced by the service option of IS-733, operating at the Rate Set 2. Rate Set 2 includes the rates 13.3 kbps, 6.2 kbps, 2.7 kbps, and 1.0 kbps. The quality of IS-733 was competitive with other wireless standards, and close to the quality of wireline telephone services. However, IS-733 operates on a higher ABR than IS-96C, resulting in reduced system capacity. The next speech coding option, IS-127 (known as EVRC), was designed to solve this problem [1]. IS-127 operates at Rate Set 1, similar to IS-96C, but delivers a quality that is equivalent to IS-733. Despite these advantages, IS-127 does not make a full use of the Rate Set 1, since it does not use the quarter rate at all, and the half rate is used only for about 5% of the frames of active speech. Moreover, although IS-127 is a variable-rate coder, its rate determination algorithm (RDA) is non-flexible and does not provide the option to operate at various ABRs. Different ABRs can be used, for example, to improve the capacity of high-traffic wireless cells, by reducing the ABR, or to increase the speech quality in low-traffic cells, by increasing the ABR. Different ABRs might also be used as pricing options by wireless service providers.

To improve the speech coding efficiency and the CDMA system flexibility, the TIA, and later the 3GPP2, managed a competition and a selection process for a new speech coding system for CDMA applications. The new system, which was designed to operate of Rate Set 1, similar to IS-96C and IS-127, is called the Selectable Mode Vocoder (SMV). The SMV name signifies the ability of the new speech coding system to operate at different operation modes, each with a different ABR, providing the flexibility for a tradeoff between the speech quality and the CDMA system capacity.

Three modes of operation were defined for the SMV selection phase, although it was envisioned the SMV system will be able to operate at each point on a smooth curve of quality vs. ABR. The three modes of operations were defined as Mode 0, or premium mode, Mode 1, or standard mode, and Mode 2, or economy mode. Table 1 gives the rate usage percentages and the ABRs for IS-127 and for the 3 modes of Conexant's SMV system. The data was measured for single coding on the clean speech database used in the selection test. The voice activity in this database is about 75%, which is non-typical for conversational speech. For conversational speech, which has about 45% voice activity, the percentage of the eighth-rate usage would increase and the ABR would decrease. For conversational speech, the usage percentages of the full-, half-, and quarter-rate usage would drop, although the relative ratios between them would be similar.

	IS-127	Mode 0	Mode 1	Mode 2
Full	71.85%	69.49%	33.06%	12.02%
Half	3.84%	6.17%	27.37%	47.52%
Quarter	0.0%	0.0%	11.57%	11.03%
Eighth	24.31%	24.35%	28.01%	29.43%
ABR	7.4 kbps	7.3 kbps	5.1 kbps	4.1 kbps

Table 1: Rate usage percentages and ABR for IS-127 and of Conexant's SMV system

The quality requirements for the SMV's 3 modes of operation was loosely defined, but Mode 0 was expected to perform better than IS-127, Mode 1 to perform similar to IS-127, and Mode 2 to perform not much worse than IS-127.

This paper describes Conexant's candidate for the SMV selection phase. Section 2 is an outline of Conexant SMV algorithm, Section 3 is a review of the eX-CELP technology used in the SMV algorithm, and Section 4 gives the details of the speech coding and parameter quantization. Section 5 reviews the results of the SMV selection test, and Section 6 concludes this paper.

2. OUTLINE OF CONEXANT SMV ALGORITHM

The Conexant SMV candidate is based on 4 codecs, a full-rate codec at the rate of 8.5 kbps, a half-rate codec at the rate of 4.0 kbps, a quarter-rate codec at the rate of 2.0 kbps, and an eighth-rate codec at the rate of 0.8 kbps. The same 4 codecs are used for all of the SMV modes, with the exception of the quarter-rate codec, which is not allowed in Mode 0 (for compatibility with IS-127). The different ABRs are achieved by an RDA that decides on the appropriate coding rate for each frame, based on the controlling mode of operation. For example, a frame of stationary voiced speech would be coded using the full-rate codec when the SMV system is operated in Mode 0, but might be coded with the half-rate codec when the SMV system is operated in Mode 1 or in Mode 2.

Figure 1 is a block diagram of the Conexant SMV system. The pre-processing includes standard high-pass filtering, noise suppression similar to IS-127, and adaptive tilt compensation. The frame processing includes LPC analysis, open-loop pitch search, signal modification, and classification. For each frame, the RDA selects one of the 4 possible coding rates, based on the SMV mode. The SMV system in general, and the full-rate codec and the half rate codec in particular, are based on Conexant eXtended CELP (eX-CELP) technology [2], which is outlined in Section 3. The quarter-rate codec and the eighth-rate codec, designed for the representation of stationary unvoiced speech or background noise, are based on spectrum- and energy-modulated random noise models.

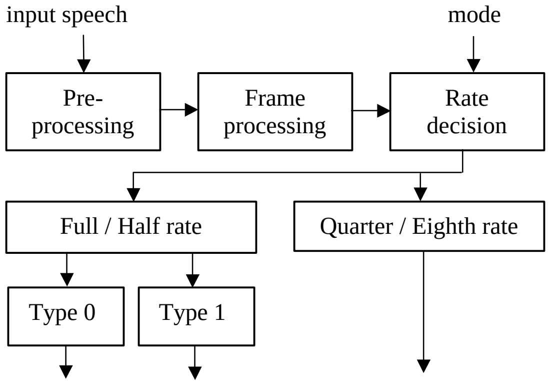

Figure 1: A block diagram of Conexant SMV candidate.

3. EX-CELP CODING

The full-rate and half-rate codecs in the Conexant SMV candidate are based on the eX-CELP approach, which was also used successfully in other speech coding systems [3]. We use the term eX-CELP to include several techniques for high quality speech compression, applicable for both low-rate and high-rate speech coding. The main theme of the eX-CELP technology is the combination of the analysis-by-synthesis (closed-loop) search approach, used extensively in traditional CELP coding, with perceptually-based decisions (open-loop) for improved perceptual representation of the speech signal. The combination of the closed-loop and the open-loop approaches, which we call closed-loop-open-loop-analysis, or COLA, is based on elaborate speech classification and parameter estimation, and is used in practically each aspect of the coding algorithm, such as signal modification, pitch search, gain quantization, and codebook search.

The signal modification scheme, which provides a stable pitch and an increased pitch prediction gain, is an integral part of the eX-CELP approach. A similar concept, sometimes called RCELP [4], was used, for example, in IS-127. Our improved eX-CELP signal modification scheme is performed on the weighted speech, using a controlled continuous warping of the signal. This approach results in an excellent quality of the modified speech, while providing substantial increase in the pitch prediction gain.

The eX-CELP employs an elaborate speech classification, and each frame is appropriately classified as either silence/background noise, stationary unvoiced, non-stationary unvoiced, onset, non-stationary voiced, or stationary voiced. A multi-level approach is used for the classification decision, starting with a VAD, followed by several stages of classification refinements. Frames of silence/background noise or stationary unvoiced frames can be represented using a spectrum- and energy-modulated noise. According to the dictated SMV operational mode, such frames might be coded using the quarter-rate codec or the eighth-rate codec. The final decision of a stationary voiced frame is based in the pitch prediction gain obtained during signal modification. Frames with a high pitch prediction gain are declared Type 1 frames, while all other frames are declared Type 0 frames.

Type 0 frames are coded similarly to traditional CELP coding. The frame is divided to subframes (4 for full-rate coding and 2 for half-rate coding), and a closed-loop pitch (adaptive codebook) search is performed for each subframe, which is followed by a fixed-codebook. For each subframe, the pitch prediction gain and the fixed-codebook gain are jointly quantized, using a 2-dimensional vector quantization scheme.

Type 1 frames are frames of stationary voiced speech, with a high and stable pitch prediction gain and a stable pitch contour throughout the frame. Such frames are also divided into subframes (4 for full-rate coding and 3 for half-rate coding). Only a single pitch value, derived from the open-loop pitch search, is used for Type 1 frames. Using the COLA approach, the pitch contour is interpolated according to the frame pitch value, and the pitch gain is calculated in an open-loop fashion for each subframe. The pitch gains of all the subframes are jointly quantized using vector quantization (VQ) before the subframe processing is performed. Since fewer bits used for the representation of the pitch lag and the pitch gains, more bits are available for the representation of the fixed codebook excitation, in comparison to Type 0 frames. The fixed-codebook search is carried out for each subframe, using the interpolated pitch contour, the quantized pitch gains, and *unquantized* fixed codebook gains. The fixed-codebook gains of all the subframes are jointly quantized, using a VQ, after the fixed-codebook contribution is determined for the whole frame.

The fixed-codebook structure and search approach is also based on the COLA approach. The fixed codebooks (except of the codebook for the full-rate Type 1 codec) consist of several sub-codebooks, each tuned for the proper representation of a particular type of speech excitation. The sub-codebooks can consist of pulse excitation of various densities and patterns, or of random-type excitation. The selection between the sub-codebooks is performed using a combination of the analysis-by-synthesis error measure, the classification information, and additional guidance based other speech parameters, for example, the estimates of the presence and the level of background noise, or an estimate of the speech peakiness.

The closed-loop error measure for each sub-codebook is provided by an efficient iterative routine for analysis-by-synthesis search in eX-CELP coding. In this approach, pulse locations are added and switched in a high-performance low-complexity analysis-by-synthesis search algorithm.

The quantization of the gains for both Type 0 and Type 1 frames also uses the COLA approach, in combining closed loop error minimization with perceptual considerations to improve the quality of the gain quantization. This scheme is beneficial in particular for noisy speech, where the gain quantization aims to keep a smooth energy envelope contour, which is an important factor in high quality noisy speech reproduction.

4. DETAILED REVIEW OF CONEXANT SMV SYSTEM

Table 2 specifies the bit allocation for the 4 rates of Conexant's SMV system. For the full-rate codec and the half-rate codec the table specifies the bit allocation for each frame type.

For the full-rate codec and the quarter-rate codec the LSFs are quantized using a 25 bit predictive multi-stage VQ. For Type 1 frames of the full-rate codec and of the quarter-rate codec, 2 bits are used to specify a spectral interpolation contour for each subframe. For the half-rate codec the LSFs are quantized using a 21-bit switch-predictor multi-stage VQ. For the eighth-rate codec the LSFs are quantized using an 11 bit predictive multi-stage VQ.

Rate	Full		Half		Quarter	Eighth
Type	0	1	0	1		
LSFs	27	25	21	21	27	11
Energy					12	5
Mode	1	1	1	1		
Pitch	26	8	14	7		
Excitation	88	120	30	39		
Gains	28	16	14	12		
Total	170	170	80	80	39	16

Table 2: Bit allocation table for Conexant's SMV system

For the full-rate and the half-rate codec, 1 bit is used to signal frame type. For frames of Type 0, the full rate codec uses 26 bits to represent the pitch lag. The full allowable pitch range is represented for the first and the third subframes, using 8 bits each, and a limited differential pitch range representation is used for the second and the fourth subframe, with 5 bits each. For frames of Type 0 the half-rate codec uses 7 bits for each of the two subframes, each represents the full allowable pitch range for that rate.

For the full-rate Type 0 codec the fixed codebook is comprised of 3 sub-codebooks, each of 5 pulses, which are represented with 22 bits for each of the four subframes. Although the number of pulses is the same for each codebook, the pulse pattern of each sub-codebook is different, and was tuned to represent different type of excitation. For example, the pulse pattern in one sub-codebook might allow a local concentration of pulses, to capture a localized pitch event. The pulse pattern in yet another sub-codebook might be more sparse, making it suitable for the representation of a noise-like excitation. The selection between the sub-codebook is performed using the COLA approach. For each sub-codebook, the best pulse combination in determined using the analysis-by-synthesis approach (closed-loop), but the selection between the codebooks incorporates the closed-loop measure with perceptual weighting (open-loop), providing a perceptually-optimal excitation source.

A single codebook of 8 pulses, represented by 30 bits for each of the four subframes, is used for full-rate Type-1 frames. The large number of bits and pulses allow a rich excitation, which is searched using a traditional analysis-by-synthesis approach.

For the half-rate Type 0 codec, the fixed codebook is comprised of 3 sub-codebooks, which are represented with 15 bits for each of the two subframes. Two of the codebooks are pulse codebooks, one comprised of 2 pulses and the other comprises 3 pulses. The third codebook is generated from a Gaussian excitation source. A particularly fast search scheme is tailored for the Gaussian codebook and analysis-by-synthesis is used for the search of the pulse codebooks. The selection between the sub-codebook is based on the COLA approach, similar to the full-rate Type 0 codec, which takes into consideration the frame classification as well as additional speech parameters.

The fixed-codebook excitation for the half-rate Type 1 codec includes only 2 sub-codebooks, one of 2 pulses and the other of 3 pulses, represented by 13 bits for each of the three subframes. Similar to full-rate Type-0 frames and half-rate Type-0 frames, the best entry in each sub-codebook is determined by the analysis-by-synthesis approach, while the selection between the two sub-codebooks is guided by additional perceptual considerations.

For frames of Type 0 the pitch gain and the fixed-codebook gain for each subframe are jointly quantized using a 2-dimensional VQ, as discussed in Section 3. For both the full-rate codec and the half-rate codec the gains are represented with 7 bits for each subframe, which results in 28 bits per frame for the full-rate codec, and 14 bits per frame for the half-rate codec.

For frames of Type 1 the pre-calculated pitch gains (4 for full-rate coding and 3 for half-rate coding) are jointly quantized, using a VQ. Six bits are used for the VQ of the pitch gains by the full-rate codec, and 4 bits are used by the half-rate codec. This significant low bit rate for the quantization of the pitch gains for Type-1 frames is the result of the stability of the pitch gains for Type-1 frames, which is one of the criterion for declaring a specific frame as a Type 1 frame.

The fixed-codebook search for a Type 1 frame is conducted using the analysis-by-synthesis approach, but the fixed-codebook gains for each subframe are left unquantized until the excitation for all of the subframes is determined. The fixed-codebook gains are jointly quantized using a VQ. The 4 fixed-codebook gains of the full-rate codec are quantized with 10 bits, and the 3 fixed-codebook gains of the half-rate codec are quantized with 8 bits. The gain quantization for Type 0 or for Type 1 frames includes a closed-loop measure and a perceptual measure, according to the COLA approach, which provides best waveform matching for clean speech and smooth energy contour for noisy speech.

The decoding operation of the Conexant SMV system is similar to other variable-rate speech coding schemes. The bits are extracted from the bit stream according to the rate information, which is transmitted through a separate control channel. The speech parameters are used to generate the speech signal, by multiplying the fixed-codebook excitation and the pitch excitation by the appropriate gains and summing them to generate the LPC excitation. The LPC excitation is passed through the LPC synthesis filter, which uses the quantized LSFs to derive the filter coefficients. An adaptive short-term and long-term post filter reduces the level of perceived coding noise.

The SMV system is targeted for CDMA applications, which are susceptible to frame errors. A CDMA system under normal operation can experience a frame erasure rate as high as 1%, and a higher frame erasure rate is also possible. Conexant's SMV system uses an elaborate backward parameter estimation and smoothing for frame erasure concealment.

5. THE SMV SELECTION PHASE

The SMV selection phase included 3 types of experiments. Experiment 1 was designed to test the codecs for clean speech in single coding at nominal, low, and high input levels, and in tandem coding at nominal level. Experiment 2 tested the codecs for clean speech under frame error conditions. Experiment 3 tested the codecs for speech in the presence of background noise. Each of the 3 experiments was repeated for the 3 operation modes of the SMV system, resulting in 9 separate experiments. The selection test results of 2 conditions from each on the 9 experiments are presented in the following figures [5].

Figure 2 shows the selection test results for clean speech in single coding and for tandem coding. (The speech material was processed with a μ-law PCM codec only after the first coding, which explains the higher performance in tandem coding for Mode 0.) Figure 3 shows the result for clean speech with channel impairments of 1% frame erasure rate and 3% frame erasure rate. The selection test results for speech in the presence of background noise are presented in Figure 4, for 15 dB vehicle noise and for 20 dB office noise. It should be emphasized that the figures might present results from different experiments, and therefore, for the same condition, the reference codecs might have different scores.

6. CONCLUSION

This paper described the candidate that was submitted by Conexant for the selection test of the SMV system. The Conexant candidate for the SMV system is based on the eX-CELP speech coding technology and was selected as the core for the SMV system for CDMA applications. A collaborative effort, which includes other companies that participated in the SMV selection phase, is under way for further improvements, refinements, and implementation of the SMV as the future standard for 3G CDMA applications.

7. REFERENCES

[1] TIA/EIA/IS-127, "*Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems*", January 1997

[2] Y. Gao et al., "*eX-CELP: A Speech Coding Paradigm*", published in the Proceedings of ICASSP 2001.

[3] J. Thyssen et al., "*A Candidate for the ITU-T 4 kbit/s Speech Coding Standard*", published in the Proceedings of ICASSP 2001.

[4] W. B. Kleijn, R. P. Ramachandran, and P. Kroon, "*Generalized Analysis-by-Synthesis Coding and its Application to Pitch Prediction*", Proc. ICASSP, 1992, pp. I337-I340.

[5] F. Corcoran, "*SMV Selection Test – Final Host and Listening Lab Report*", 3GPP2-C11-20000821-003, August 2000

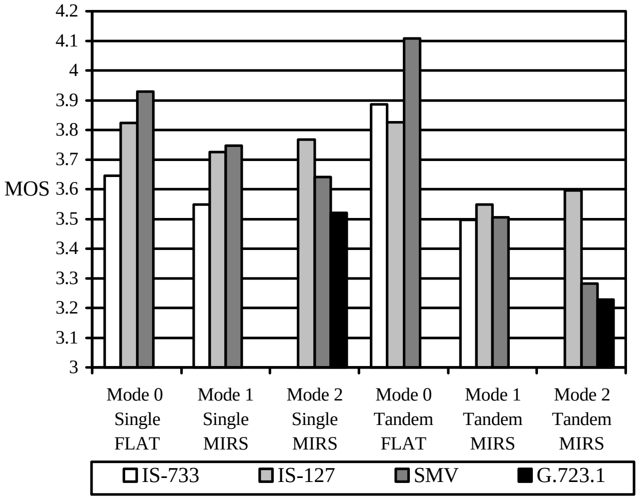

Figure 2: Results for clean speech

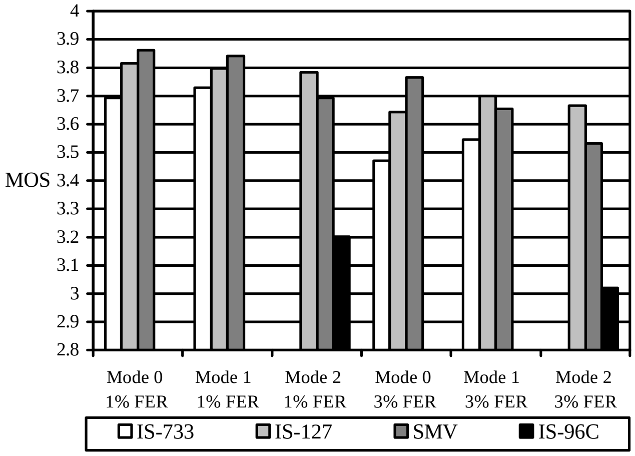

Figure 3: Results for speech in channel impairments

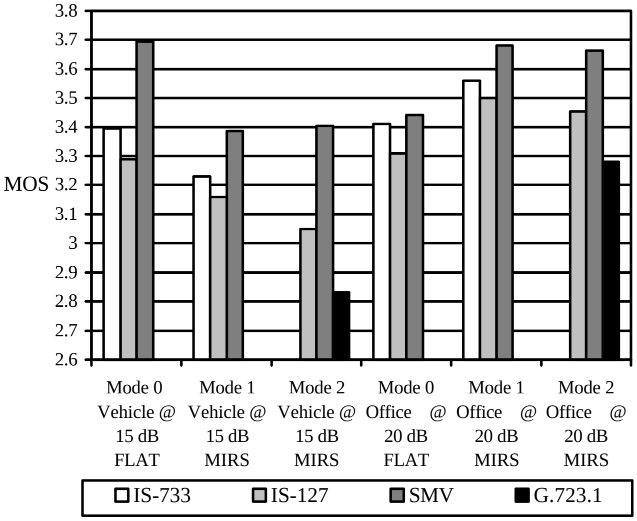

Figure 4: Results for speech in the presence of background noise