BroadVoice®16: A PacketCable Speech Coding Standard for Cable Telephony

Juin-Hwey Chen and Jes Thyssen
Broadcom Corporation
Irvine, California, USA

***Abstract* - This paper presents the BroadVoice16 (BV16) speech codec, which is a mandatory codec in the PacketCable 1.5 standard. For cable telephony based on PacketCable 1.5, BV16 possesses a set of attributes not met by other speech codecs: (1) no royalty, (2) high quality, (3) low delay, (4) low complexity, and (5) medium to low bit-rate. The royalty-free requirement excludes many modern speech coding techniques. Hence, an older paradigm is resurrected and improved as the foundation of BV16. Extensive test results including independent subjective tests, PESQ evaluation across 13 languages, and DTMF pass-through evaluation demonstrate the high performance of BV16.**

I. INTRODUCTION

In recent years, voice telephony has started the migration from conventional circuit-switched voice to Voice over Internet Protocol (VoIP) in a major way. For example, at the time of writing, Vonage® has more than 2 million VoIP subscribers, and cable operators in North America are estimated to have more than 6 million VoIP subscribers. The number of cable VoIP subscribers is currently growing at a rapid rate of approximately 11,000 subscribers per day. Such initial VoIP deployments typically use the uncompressed 64 kb/s G.711 PCM speech codec because of its simplicity and high output quality, even though it is not very efficient in bit-rate. The number of subscribers in such initial VoIP deployments has not reached a critical point for congestion to become a problem. However, with the high growth rate of subscribers, in the future the VoIP traffic is expected to cause congestion at the uplinks of cable access networks.

A simple way to prepare for the expected uplink congestion is to use a speech codec with higher compression than G.711. Ideally the best time to deploy such voice compression is not when congestion happens, but during the initial deployment phase by making end-user equipment and network equipment voice-compression-capable. That way, when congestion happens, there is no need to replace equipment, since voice compression can simply be "turned on" by software at that time.

In North America, cable telephony based on VoIP is specified in the PacketCable™ standard developed by CableLabs®, an organization jointly supported by North American cable operators to develop standards and certify equipment for cable networks. Foreseeing the need for voice compression to address uplink congestion, CableLabs selected the ITU-T G.728 and G.729 Annex E as recommended speech codecs in addition to the mandatory G.711 codec in the PacketCable 1.0 standard.

On the one hand, although the 64 kb/s G.711 codec is royalty-free, it is wasteful in transmission bit-rate. On the other hand, low-bit-rate standard codecs such as G.728, G.729E, G.729, G.723.1, and GSM-EFR are more efficient in bit-rate but carry royalties. Beside royalties, most low-bit-rate codecs have higher coding delays than desired or have high codec complexity that increases deployment costs. Some codecs also have less-than-ideal output speech quality.

The cable telephony service directly competes with the traditional Public Switched Telephone Network (PSTN) telephony service, and as such a deployed codec needs to have high speech quality and low delay. Low codec complexity and no royalty will allow cable operators to compete better in terms of costs. Furthermore, while very low bit-rate is not necessary due to the packet header overhead, medium to low bit-rate is desirable for alleviating the cable uplink congestion. Thus, for VoIP cable telephony many cable operators in North America desire to use a new speech codec with (1) no royalty, (2) high quality, (3) low delay, (4) low complexity, and (5) medium to low bit-rate.

The BroadVoice16 (BV16) [1], a 16 kb/s narrowband speech codec for 8 kHz sampling, was designed from its inception [2] in 2000 to have these five attributes above. Hence, when CableLabs announced the Request for Proposals (RFP) for a royalty-free speech codec to be mandated in the PacketCable 1.1 standard, BV16 was a natural match and was submitted as a candidate to CableLabs. Five different candidate codecs were submitted to CableLabs by the submission deadline of June 30, 2002.

The royalty-free requirement in the PacketCable 1.1 audio codec RFP excludes the possibility of using the dominant and powerful speech coding method known as Code Excited Linear Prediction (CELP) [3] and its related

techniques, since they have been extensively researched and patented. To avoid such CELP-related patents, an older paradigm known as Noise Feedback Coding (NFC) [4]-[7] was resurrected and improved as the foundation for BV16. The resulting Two-Stage Noise Feedback Coding (TSNFC) [8], [9] technique performs noise feedback coding in a novel nested structure to provide both short-term and long-term prediction as well as both long-term and short-term noise spectral shaping.

After evaluating the submitted candidate codecs, including commissioning AT&T Voice Quality Assessment Laboratory to conduct formal subjective listening tests, in April 2004 CableLabs selected BV16 as one of the two mandatory codecs for the PacketCable 1.5 standard. The Real-Time Protocol (RTP) payload format of BV16 is specified in Internet Engineering Task Force (IETF) RFC 4298 [10]. The BV16 codec was later voted as an SCTE® (Society of Cable Telecommunications Engineers) standard in May 2006, and in September 2006 it became an ANSI American national standard for Voice over IP Applications in Cable Telephony [1]. In October 2006, ITU-T Study Group 9 approved a draft revised ITU-T Recommendation J.161 which lists BV16 as a mandatory codec for IPCableCom and gives a normative reference to the ANSI/SCTE BV16 standard.

This paper presents the BroadVoice16 codec algorithm and its attributes, including the bit allocation, coding delay, complexity, and performance.

II. BROADVOICE16 CODEC ALGORITHM

Conventional NFC codecs [6], [7] use scalar quantization to quantize the prediction residual, and usually the NFC principle is applied only to the short-term prediction and short-term noise spectral shaping. Some NFC codecs may add long-term prediction [7] or long-term noise spectral shaping [11], but not both. The BroadVoice16 codec improves these earlier NFC codecs by (i) adding both long-term prediction and long-term noise spectral shaping in a novel two-stage nested structure of TSNFC, and (ii) replacing scalar quantization by vector quantization (VQ) for the prediction residual [8], [9]. A high-level description of the BV16 coding algorithm is given in this section.

A. *Codec Structure*

The encoder structure of BV16 is shown in Fig. 1. It is based on the TSNFC Form 3 structure described in [8]. The BV16 decoder structure is shown in Fig. 2. An optional postfilter can be added to the BV16 decoder.

To achieve a very low coding delay, BV16 uses a small frame size of 5 ms (40 samples at 8 kHz sampling) and no look-ahead. The resulting algorithmic buffering delay of 5 ms is much lower than the 15 to 40 ms algorithmic delay of most other ITU-T low-bit-rate codecs. The parameters of the short-term and long-term predictors and the excitation gain are transmitted once a frame.

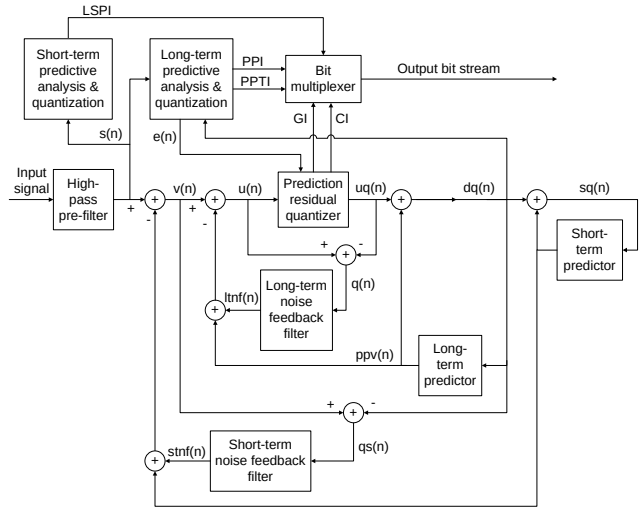

Fig. 1 High-level block diagram of the BV16 encoder

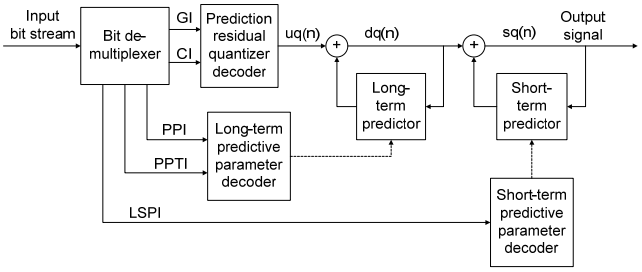

Fig. 2 High-level block diagram of the BV16 decoder

B. *Short-Term Prediction and Short-Term Noise Spectral Shaping*

BV16 uses a short-term predictor order of 8. The autocorrelation method of LPC analysis is used with a 20 ms asymmetrical window, with the peak of the window centered at the current 5 ms window. The short-term predictor coefficients are converted to the Line-Spectrum Pair (LSP) parameters, which are then quantized by an 8^{th}-order moving average (MA) predictive coding scheme with two-stage VQ of the LSP inter-frame prediction residual. The first stage employs a 7-bit, 8-dimensional VQ with unconstrained codebook, and the codebook search uses the mean squared error (MSE) distortion measure. The second stage uses a 7-bit, 8-dimensional sign-shape VQ and a weighted mean squared error (WMSE) distortion measure.

To make the LSP quantizer more robust to bit errors, special handling is used in both the LSP encoder and LSP decoder to detect bit errors without sending redundant bits. In the second-stage LSP VQ, only codevectors that preserve

the ordering for the first three LSP parameters are considered in the codebook search. A violation of the ordering property at the decoder indicates a transmission error, and in this case the LSP parameters of the last frame are used. This constrained LSP VQ scheme causes practically no degradation in clean channel conditions.

Let a_i, i = 1, 2, …, 8 be the coefficients of the unquantized short-term prediction error filter $A(z)$. BV16 uses a short-term noise feedback filter in the form [9] of

$$F_s(z) = \frac{\sum_{i=1}^{8} a_i \cdot \left(\gamma_1^{\,i} - \gamma_2^{\,i}\right) \cdot z^{-i}}{\sum_{i=0}^{8} a_i \cdot \gamma_2^{\,i} \cdot z^{-i}}, \quad 0 < \gamma_1 < \gamma_2 < 1,$$

which produces a noise spectral envelope given by

$$N_s(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)},$$

where $\gamma_1 = 0.5$ and $\gamma_2 = 0.85$.

C. Long-Term Prediction and Long-Term Noise Spectral Shaping

BV16 uses a three-tap long-term predictor with an integer pitch period. To keep the codec complexity low, both the pitch period and the predictor taps are determined in an open-loop fashion. The pitch period is extracted using a 4:1 decimated 2 kHz signal in the first stage and pitch refinement in the second stage. The pitch period is represented by 7 bits, with a range from 10 to 136 samples. The three pitch taps are vector quantized to 5 bits, with the energy of the open-loop pitch prediction residual used as the distortion measure for the codebook search.

When the input speech is voiced, BV16 uses long-term noise spectral shaping to shape the coding noise spectrum according to the harmonic structure of the voiced speech spectrum. This is achieved through the long-term noise feedback filter in Fig. 1. To keep the complexity low, this filter is chosen to have the simple form of

$$F_l(z) = N_l(z) - 1 = \lambda\, z^{-pp},$$

where pp is the pitch period, and λ is proportional to the optimal tap weight of a single-tap pitch predictor.

D. Gain Quantization

To keep the complexity low, BV16 also determines the excitation gain in an open-loop fashion. The average power of the open-loop prediction residual signal (after short-term and long-term prediction) within the current frame is calculated and converted to the base-2 logarithmic domain. The resulting log-gain is then quantized using 8th-order inter-frame MA predictive coding. The inter-frame log-gain prediction error is scalar quantized to 4 bits.

The log-gain quantizer also imposes a special constraint on how fast the log-gain can increase given the log-gain value and the log-gain increase in the previous frame. At the decoder, if such a constraint is violated, it indicates a transmission error, and in this case the log-gain of the last frame is used. This scheme detects certain gain bit errors without sending redundant bits.

E. Excitation Vector Quantization

The excitation signal is quantized with 4-dimensional VQ. The excitation VQ codebook has a sign-shape structure, with 1 bit for sign and 4 bits for shape. This gives an effective excitation encoding bit-rates of 1.25 bits/sample.

The analysis-by-synthesis principle is used in the excitation VQ codebook search. Conceptually, each of the excitation VQ codevector is scaled and passed through the feedback filter structure in Fig. 1, and the one that minimizes the energy of the quantization error vector $q(n)$ is selected [1]. As shown in [8], the codebook search complexity can be greatly reduced by decomposing the quantization error vector $q(n)$ into the zero-input response (ZIR) and zero-state response (ZSR). Further complexity reduction can be achieved using the techniques proposed in [9]. For an explanation of the basic concepts and detailed excitation VQ codebook search procedure in TSNFC, see [2], [8], and [9].

F. Bit Allocation

Table 1 shows the BV16 bit allocation for each 5 ms frame. The side information takes 30 bits/frame and the excitation vectors take 50 bits/frame.

TABLE 1 BV16 BIT ALLOCATION

Parameter	Number of bits
LSP	7+7=14
Pitch period	7
3 pitch taps	5
Excitation gain(s)	4
Excitation vectors	(1+4)×10=50
Total per frame	80

G. Example Adaptive Postfilter and Example Packet Loss Concealment

In the ANSI BV16 standard, the adaptive postfilter (PF) and packet loss concealment (PLC) algorithms are optional algorithmic components. An example PF and an example PLC are specified in the ANSI BV16 specification [1]. However, other PF and PLC algorithms can also be used with BV16, since PF and PLC are post-processing steps and do not affect bit-stream compatibility.

III. BroadVoice16 Codec Complexity

A. Computational Complexity

Most CELP-based narrowband standard codecs, such as G.728, G.729, G.729E, G.723.1, EVRC, AMR, etc., have a computational complexity between 20 and 36 MIPS on a 16-bit fixed-point DSP. With emphasis in low-complexity design, BV16 has a considerably lower complexity than most other comparable codecs. Specifically, an efficient implementation of BV16 only takes about 10 MIPS for the encoder and 2 MIPS for the decoder, resulting in 12 MIPS for a full-duplex channel. This complexity is about a factor of 2 to 3 lower than the other codecs mentioned above.

B. Memory Size Requirement

BV16 requires lower memory sizes than most other comparable codecs. While most CELP-based narrowband standard codecs require 2.5 to 4.6 kwords of RAM, BV16 requires about 2 kwords of RAM. Similarly, for the total memory footprint (including RAM, data tables, and program size), most CELP-based standard codecs require more than 20 kwords, but BV16 requires only about 13 kwords.

IV. BroadVoice16 Codec Performance

A. PESQ Evaluation

The ITU-T P.862 Perceptual Evaluation of Speech Quality (PESQ) was used to evaluate the objective speech quality of BV16 and many other standard codecs. To assess the robustness of BV16 and other standard codecs across different languages, a large scale PESQ evaluation was conducted with 13 different languages using the speech materials in Arabic, Chinese, English, French, German, Hindi, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, and Thai from the NTT-AT 1994 multi-lingual speech database. For each codec evaluated, the PESQ score for each of the 96 sentence pairs (8 seconds each) from each of the 13 languages are computed individually and the resulting 96 × 13 = 1248 PESQ scores are averaged. This process closely resembles how the Mean Opinion Score (MOS) in a subjective listening test is calculated. The resulting averaged PESQ scores are summarized in Table 2.

It can be seen that BV16 is ranked higher than all other low-bit-rate standard codecs listed in the table in terms of the average PESQ across the 13 languages. It is only second to the 64 kb/s G.711 μ-law codec.

The PESQ score averaged over each language and each codec is shown in Fig. 3. The PESQ scores of BV16 are relatively consistent across the 13 languages, ranging from about 4.04 to 4.12, with a range of only 0.08. All other codecs in the table have larger ranges of PESQ scores, with some of them having PESQ ranges approaching 0.2.

TABLE 2 PESQ Averaged over 13 Languages

Codec	PESQ
G.711 μ-law	4.119
BV16	4.077
G.729E	4.040
GSM-EFR	4.008
G.726 at 32 kb/s	3.938
G.728	3.891
G.729	3.798
G.723.1 at 6.3 kb/s	3.629
G.729D	3.557
G.723.1 at 5.3 kb/s	3.489

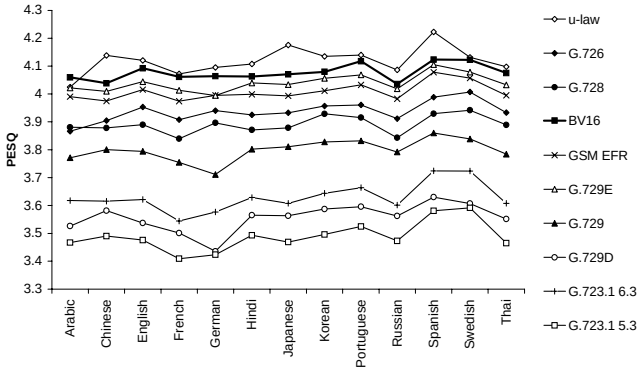

Fig. 3 Narrowband PESQ across 13 languages

B. Subjective Listening Test Results

Dynastat, Inc. conducted a formal subjective listening test to compare BV16 with other ITU-T standard codecs. A total of 32 naïve listeners participated in this MOS listening test. Table 3 shows the MOS scores of BV16 and some other common ITU-T toll-quality codecs. Statistical analysis showed that BV16 was rated statistically better than toll-quality codecs G.729, G.726 at 32 kb/s, and G.728 in this test. This confirms that BV16 is a toll-quality speech codec.

TABLE 3 Mean Opinion Scores from the Dynastat Test

Codec	MOS
G.711 μ-law	3.91
BV16	3.76
G.729	3.56
G.726 at 32 kb/s	3.56
G.728	3.54

C. DTMF Tones Pass-Through Test Results

It is desirable for a speech codec to be able to encode and decode the Dual-Tone Multi-Frequency (DTMF) signaling tones without affecting the subsequent DTMF detection accuracy. The performance of BV16 in passing real-world DTMF tones in-band was measured and compared against that of the ITU-T G.711, G.728, G.729, and G.729E codecs. The main metric used was the DTMF detection error rate after the DTMF signal was encoded and

decoded by the voice codec under test and subsequently passed to a DTMF detector that is widely used in commercial VoIP products.

To find the typical DTMF tone durations encountered in the real world, numerous DTMF tones generated by humans and automatic dialers were recorded and analyzed. It was found that the typical DTMF tone durations were in the range of 90 to 160 ms. Even when a person deliberately touched the telephone keypad as briefly as possible, the DTMF tone duration was never found to be shorter than 60 ms. Thus, 60 ms was used as the minimum tone duration in this test designed to evaluate the performance of different codecs in passing real-world DTMF signals.

Using 60 ms tone durations, 60 ms gaps between tones bursts, -4 dBm DTMF signal level, and a typical SNR of 36 dB as the base line, we tested the five codecs using the 24 conditions listed in Table 4. A DTMF simulator was used to generate a string of 20,000 random DTMF digits for each of the 24 conditions. Each string of 20,000 DTMF digits were then encoded and decoded by each of the five codecs and then passed through a commercial DTMF detector. The DTMF detector output digit string is compared with the original input digit string to see if there is any digit error.

The resulting DTMF detection error rates are shown in Table 4. The results showed that BV16, G.711, and G.728 each passed all of the nearly half a million DTMF digits without causing a single DTMF detection error. G.729E caused DTMF detection errors in 8 of these 24 test conditions. Due to its lower bit-rate presumably, G.729 caused detection errors in 21 test conditions, with the error rates in some conditions reaching as high as 20 to 29%.

TABLE 4 DETECTION ERROR RATES FOR DTMF PASS-THROUGH

Condition		G.711	G.728	BV16	G.729E	G.729
Balanced Pair	Normal	0.00%	0.00%	0.00%	0.00%	1.75%
	-14 dBm tone level	0.00%	0.00%	0.00%	0.00%	1.71%
	-9 dBm tone level	0.00%	0.00%	0.00%	0.00%	1.84%
	75 ms tone duration	0.00%	0.00%	0.00%	0.00%	0.13%
	90 ms tone duration	0.00%	0.00%	0.00%	0.00%	0.02%
	105 ms tone duration	0.00%	0.00%	0.00%	0.00%	0.00%
	SNR 10 dB higher	0.00%	0.00%	0.00%	0.00%	3.49%
	SNR 10 dB lower	0.00%	0.00%	0.00%	0.00%	1.84%
High Tone 6 dB lower	Normal	0.00%	0.00%	0.00%	0.00%	1.28%
	-14 dBm tone level	0.00%	0.00%	0.00%	0.01%	1.19%
	-9 dBm tone level	0.00%	0.00%	0.00%	0.01%	1.20%
	75 ms tone duration	0.00%	0.00%	0.00%	0.00%	0.05%
	90 ms tone duration	0.00%	0.00%	0.00%	0.00%	0.00%
	105 ms tone duration	0.00%	0.00%	0.00%	0.00%	0.00%
	SNR 10 dB higher	0.00%	0.00%	0.00%	0.00%	1.70%
	SNR 10 dB lower	0.00%	0.00%	0.00%	0.01%	2.10%
High Tone 3 dB higher	Normal	0.00%	0.00%	0.00%	0.02%	21.09%
	-14 dBm tone level	0.00%	0.00%	0.00%	0.04%	19.79%
	-9 dBm tone level	0.00%	0.00%	0.00%	0.02%	20.17%
	75 ms tone duration	0.00%	0.00%	0.00%	0.00%	10.15%
	90 ms tone duration	0.00%	0.00%	0.00%	0.00%	9.33%
	105 ms tone duration	0.00%	0.00%	0.00%	0.00%	5.61%
	SNR 10 dB higher	0.00%	0.00%	0.00%	0.09%	28.74%
	SNR 10 dB lower	0.00%	0.00%	0.00%	0.01%	13.64%

V. CONCLUSION

This paper presented the motivation for developing BV16, how BV16 fills a void that no other standard speech codec fills, and how the attributes of BV16 align with VoIP cable telephony. The paper also outlined the structure and algorithm of BV16 and presented an overview of the low complexity and high performance that led to the standardization of BV16 by CableLabs, SCTE, and ANSI. In summary, BV16 is a medium to low bit-rate codec with high quality, low delay, low complexity, and no royalty for PacketCable 1.5 applications.

ACKNOWLEDGMENTS

We would like to thank Cheng-Chieh Lee and Robert Zopf for their partial contributions in the floating-point C codes, fixed-point C codes, optimized assembly codes, and performance testing of the BroadVoice16 codec.

REFERENCES

[1] "BV16 speech codec specification for voice over IP applications in cable telephony," American National Standard, ANSI/SCTE 24-21 2006.

[2] J.-H. Chen, US Patent Application No. 20000722077, "Method and apparatus for one-stage and two-stage noise feedback coding of speech and audio signals," filed November 2000.

[3] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): high quality speech at very low bit rates," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 937 - 940, March 1985.

[4] C. C. Cutler, US Patent No. 2,927,962, "Transmission systems employing quantization," filed April 1954, issued March 1960.

[5] E. G. Kimme and F. F. Kuo, "Synthesis of optimal filters for a feedback quantization system," *IEEE Trans. Circuit Theory*, pp. 405-413, September 1963.

[6] J. D. Makhoul and M. Berouti, "Adaptive noise spectral shaping and entropy coding in predictive coding of speech," *IEEE Trans. Acoust., Speech, Sig. Proc.*, pp.63-73, February 1979.

[7] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Sig. Proc.*, pp. 247-254, June 1979.

[8] J.-H. Chen, "Novel codec structures for noise feedback coding of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. I-681 - I-684, May 2006.

[9] J. Thyssen and J.-H. Chen, "Efficient VQ techniques and general noise shaping for noise feedback coding," *Proc. Interspeech 2006 ICSLP*, pp. 221 - 224, September 2006.

[10] J.-H. Chen, W. Lee, and J. Thyssen, "RTP payload format for BroadVoice speech codecs," IETF RFC 4298, December 2005.

[11] C. C. Lee, "An enhanced ADPCM coder for voice over packet networks," *Int'l J. Speech Tech.*, pp. 343-357, May 1999.