THE BROADVOICE SPEECH CODING ALGORITHM

Juin-Hwey Chen and Jes Thyssen

Broadcom Corporation, Irvine, California, USA

ABSTRACT

This paper describes the BroadVoice® speech coding algorithm, which has been standardized as PacketCable™, SCTE®, and ANSI standards for VoIP cable telephony. The BroadVoice family of codecs includes a 16 kb/s BroadVoice16 narrowband codec and a 32 kb/s BroadVoice32 wideband codec. BroadVoice is based on two-stage noise feedback coding with vector quantization of excitation and is optimized for low delay, low complexity, and high quality. It has an algorithmic delay of merely 5 ms and a complexity of only 12 and 17 MIPS for BroadVoice16 and BroadVoice32, respectively. The speech quality of BroadVoice16 is better than that of G.728, G.729, and 32 kb/s G.726. The speech quality of BroadVoice32 is better than that of 64 kb/s G.722.

Index Terms — BroadVoice, BV16, BV32, NFC, noise feedback coding.

1. INTRODUCTION

In 2002, CableLabs®, the standardization arm of cable operators in North America, set out to establish a new speech coding standard for VoIP cable telephony. One of the speech codecs CableLabs standardized for PacketCable 1.5 was BroadVoice16 (or BV16 for short) [1], which is a 16 kb/s narrowband codec for 8 kHz sampling. BV16 was later voted as an SCTE (Society of Cable Telecommunications Engineers) standard in May 2006 and became an ANSI American National Standard in September 2006 [2].

CableLabs is currently standardizing PacketCable 2.0 for next-generation VoIP cable telephony. The PacketCable 2.0 codec and media specification specifies all codecs as optional. BV16 is listed as one of the codecs in the PacketCable 2.0 standard. BroadVoice32 (or BV32), which is a 32 kb/s wideband codec for 16 kHz sampling and is similar to BV16 in algorithm, is also listed as one of the wideband codecs in the PacketCable 2.0 standard.

Unlike most other modern speech coding standards based on Code Excited Linear Prediction (CELP) [3], BV16 and BV32 are not CELP coders. Instead, they are based on two-stage noise feedback coding (TSNFC) [4], [5]. The BroadVoice codecs were designed from the start to be optimized for VoIP applications. The main design goal of BroadVoice was to make the coding delay and codec complexity as low as possible while maintaining output speech quality as close to transparent as possible.

To achieve a low coding delay, both BV16 and BV32 use a frame size of 5 ms and do not use any look ahead. Hence, the total algorithmic buffering delay of BroadVoice is merely 5 ms. With the exception of ITU-T G.728 Low-Delay CELP, this 5 ms algorithmic delay is lower than that of all other modern speech coding standards, which typically have algorithmic delays in the range of 15 to 40 ms. In addition, to achieve low codec complexity, efficient vector quantization (VQ) techniques [5], [6] were developed for BroadVoice, and the rest of the codec design strived to minimize the codec complexity as much as possible.

This paper is organized as follows. Sections 2 to 8 give a high-level description of the BV16 and BV32 algorithms. Sections 9 and 10 discuss the codec complexity and performance, respectively, and Section 11 gives some concluding remarks.

2. BASIC CODEC STRUCTURES

Figures 1 and 2 show the high-level block diagrams of the BV32 encoder and BV32 decoder, respectively. The BV32 encoder is based on the TSNFC Form 2 structure described in [5]. The input signal is first passed through a fixed high-pass pre-filter to remove possible DC bias or low frequency rumble. The filtered signal is further passed through a fixed pre-emphasis filter that gives a high-pass spectral tilt. The resulting pre-emphasized signal is then encoded using the TSNFC Form 2 structure.

The BV32 encoder encodes the short-term predictor parameters, long-term predictor parameters, excitation gain, and excitation vectors. The BV32 decoder decodes these parameters and passes a scaled excitation signal through the long-term synthesis filter and short-term synthesis filter to get a quantized version of the pre-emphasized signal, which is then passed through a de-emphasis filter to get the BV32 decoder output signal.

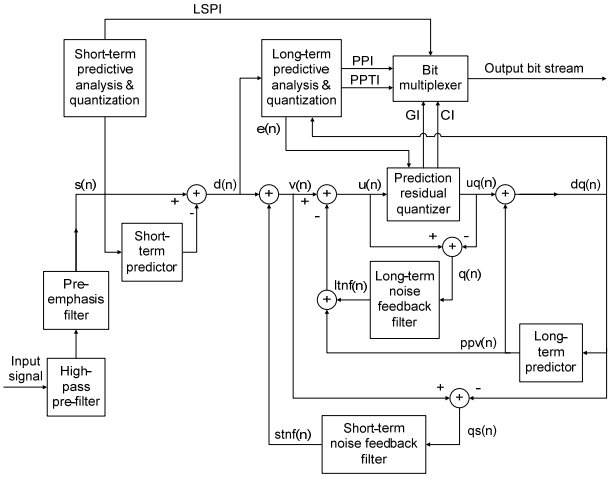

Fig. 1 High-level block diagram of the BV32 encoder

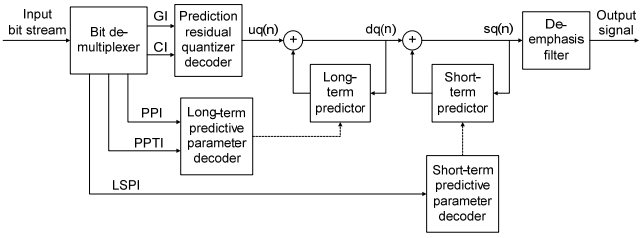

Fig. 2 High-level block diagram of the BV32 decoder

Figure 3 shows the high-level block diagram of the BV16 encoder. The BV16 encoder is based on the TSNFC Form 3 structure described in [5]. The main difference from the BV32 encoder is that the short-term predictor and short-term noise feedback filter loops have changed, and that the pre-emphasis filter is not used in BV16. The rest of the coding algorithm is very similar to BV32. The BV16 decoder is the same as the BV32 decoder in Fig. 2 except that there is no de-emphasis filter. An optional postfilter can be added to the BV16 decoder.

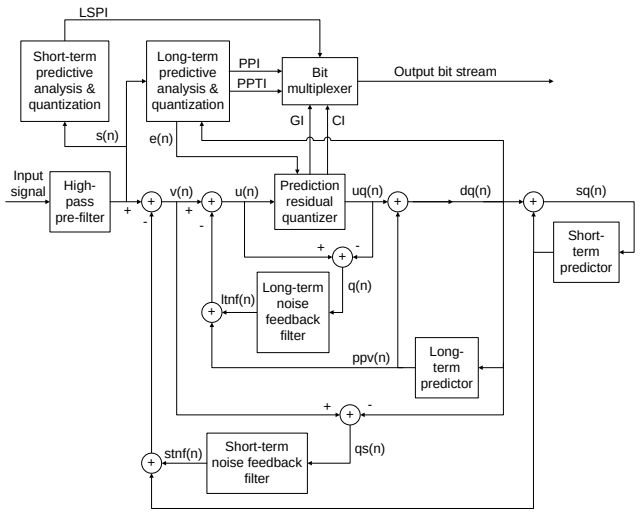

Fig. 3 High-level block diagram of the BV16 encoder

3. SHORT-TERM PREDICTION AND SHORT-TERM NOISE SPECTRAL SHAPING

To keep the complexity low, BV16 and BV32 use a low short-term predictor order of 8, and the LPC analysis window length is fixed at 160 samples (20 ms for BV16 but only 10 ms for BV32). The analysis window is asymmetric, with the peak of the window located at the center of the current 5 ms frame. Autocorrelation LPC analysis based on Levinson-Durbin recursion is used to derive the 8th-order LPC predictor coefficients, which are converted to the Line-Spectrum Pair (LSP) parameters [7].

Due to the small 5 ms frame size, it is necessary to use inter-frame predictive coding for the LSP and gain; otherwise, the side information bit rate will be too high. The LSP parameters are quantized using a fixed 8th-order moving-average (MA) predictor that covers a time span of 8 × 5 ms = 40 ms. The inter-frame LSP prediction residual is quantized using two-stage VQ. The first stage uses 8-dimensional VQ with a 7-bit codebook. For BV32, the second stage uses split VQ with a 3-5 split and 5 bits each. For BV16, the second stage is full 8-dimensional VQ with 1-bit sign and 6-bit shape. Both BV16 and BV32 use the mean squared error (MSE) distance measure for the first stage and weighted mean squared error (WMSE) for the second stage.

Although BroadVoice was mainly optimized for VoIP, it was also designed to be reasonably robust to bit errors for other applications where bit errors could be an issue.

To make the decoded LSP parameters more robust to bit errors, the LSP encoder and decoder both use special handling to detect bit errors without sending redundant bits. During second-stage VQ, only codevectors that preserve the ordering for the first three LSP parameters are considered. A violation at the decoder indicates a transmission error, and error concealment in the form of LSP repetition is used. This constraint of order preservation was a compromise between error detection and clear channel speech quality, favoring minimal degradation to clear channel speech quality. Since the constraint enforced at the encoder is consistent with the ordering property of LSP parameters for stable filters [7], practically no degradation is observed in clean channel conditions.

In terms of TSNFC structures, Form 2 of BV32 is generally of lower complexity than Form 3 of BV16, but Form 3 has more flexibility in shaping the spectral envelope of the coding noise [5]. With the 8th-order short-term prediction error filter given by

$$A(z) = \sum_{i=0}^{8} a_i z^{-i} ,$$

BV32 uses a short-term noise feedback filter $F(z) = 1 - \tilde{A}(z/\gamma)$, $0 < \gamma < 1$, resulting in a noise spectral envelope shape given by

$$N_{BV32}(z) = \frac{\tilde{A}(z/\gamma)}{\tilde{A}(z)} ,$$

where the "tilde" of $\tilde{A}(z)$ indicates usage of quantized short-term predictor coefficients. This noise shape is tied directly to the quantized short-term predictor, and hence only the numerator polynomial can be controlled through γ. The Form 3 structure of BV16 uses a short-term noise feedback filter in the form [6] of

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

producing a coding noise with a spectral envelope given by

$$N_{BV16}(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} .$$

Since the quantized short-term predictor is not part of this, it is advantageous to use un-quantized short-term predictor coefficients. BV32 uses $\gamma = 0.75$, while BV16 uses $\gamma_1 = 0.5$ and $\gamma_2 = 0.85$.

4. LONG-TERM PREDICTION AND LONG-TERM NOISE SPECTRAL SHAPING

For long-term prediction, a three-tap pitch predictor with an integer pitch period is used. To keep the complexity low, the pitch period and the pitch taps are both determined in an open-loop fashion, i.e., no closed-loop search is performed. The integer pitch period is represented by 7 bits for BV16 and 8 bits for BV32.

The pitch period is extracted using a decimation-based approach to reduce the complexity. The short-term prediction residual is passed through a weighted short-term synthesis filter to get a weighted input signal. This weighted signal is down-sampled to 2 kHz with 4:1 and 8:1 decimation for BV16 and BV32, respectively. A coarse pitch period is extracted using this 2 kHz weighted signal by identifying the time lag that gives the maximum energy-normalized correlation subject to some decision logic to eliminate integer multiples of the true pitch period. The resulting coarse pitch period is then converted to the undecimated domain, and a pitch refinement search is performed around the converted coarse pitch period using the undecimated weighted input signal. The resulting pitch period is the final pitch period.

To capture the periodicity in the waveforms of high-pitched musical instruments such as trumpet and saxophone, the minimum pitch period is chosen to be 10 samples for BV16 and BV32. The maximum pitch period is 136 and 264 for BV16 and BV32, respectively. The remaining pitch period of 137 and 265 are reserved for possible system signaling to the speech decoder.

The three pitch predictor taps are jointly quantized using a 5-bit vector quantizer for both BV16 and BV32. The distortion measure used in the pitch tap codebook search is the energy of the open-loop pitch prediction residual. The 32 codevectors in the pitch tap codebook have been "stabilized" [8] to make sure that they will not give rise to an unstable long-term synthesis filter.

When the input speech is a quasi-periodic voiced signal, the long-term noise shaping in BroadVoice shapes the noise spectrum to follow the harmonic structure of the voiced speech spectrum to some degree. This is achieved through the long-term noise feedback filter in the feedback loop of Figs. 1 and 3. 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 extracted above,

$$\lambda = \begin{cases} 0.5, & \beta \geq 1 \\ 0.5\beta, & 0 < \beta < 1 \\ 0, & \beta \leq 0 \end{cases}$$

and β is the optimal tap weight of a single-tap pitch predictor for the weighted input speech signal using the pitch period of pp. The noise spectral shape after such long-term noise spectral shaping is given by the frequency response of the filter represented by $N_l(z)$.

5. GAIN QUANTIZATION

The excitation gain is also determined in an open-loop fashion to keep the complexity low. For BV16, a single excitation gain is transmitted for each 5 ms frame. For BV32, two excitation gains are transmitted, with each gain corresponding to a 2.5 ms sub-frame. The prediction residual signal after open-loop short-term and long-term prediction is first calculated. The average power of the prediction residual signal within the current frame or sub-frame is then calculated and converted to the base-2 logarithmic domain. The resulting log-gain is then quantized using inter-frame or inter-sub-frame MA predictive coding. The MA predictor order for the log-gain is 8 for BV16 and 16 for BV32, with both covering a time span of 40 ms. The MA predictor coefficients are fixed. Each log-gain prediction residual sample is quantized to 4 bits for BV16 and 5 bits for BV32 using scalar quantization.

Bit errors can cause the decoded gain to be much larger than the transmitted gain, resulting in big audible pops in the decoded speech. Such distortion can be avoided by a "gain-change limitation" scheme, which places a dynamic constraint on the maximum increase of log-gain that is allowed at the gain encoder and gain decoder. Specifically, a constraint matrix is trained off-line using hours of natural speech. The row index of the matrix points to different bins of log-gain relative to a long-term average log-gain. The column index points to different bins of the log-gain change between adjacent frames (in BV16) or sub-frames (in BV32). The value of each matrix element is the largest log-gain change allowed during gain encoding and decoding, given that the log-gain and log-gain change of the last frame or sub-frame is in the bin defined by the row index and column index of that matrix element. The matrix elements can be chosen to be a certain percentile (e.g. 99.9%) of the actual observed log-gain change values during BroadVoice encoding of a long training speech file.

During gain encoding, the gain encoder first performs normal predictive quantization of the log-gain. The last set of the quantized log-gain and log-gain change in the last frame or sub-frame is used to derive the row and column indices to the constraint matrix. The corresponding matrix element is used as a threshold. If the quantized log-gain represents a gain increase from last frame or sub-frame smaller than the threshold, this quantized log-gain is accepted and the corresponding gain index transmitted; otherwise, the next largest quantized log-gain is considered. This process continues until the largest quantized log-gain that satisfies the gain-change limitation constraint is found. The resulting log-gain is accepted as final quantized log-gain, and the corresponding gain index is transmitted. In the rare occasion when all the possible quantized log-gain values do not satisfy the gain-change limitation threshold, the smallest of them is accepted.

During gain decoding, a similar procedure is used to extract the threshold from the constraint matrix. If the normally decoded log-gain gives rise to a gain increase smaller than the extracted threshold, or if it corresponds to the smallest gain in the gain codebook, it is accepted as the decoded log-gain. Otherwise, the decoded log-gain has been corrupted by bit errors and the decoded log-gain of the last frame or sub-frame is used instead.

As described, this gain-change limitation scheme can detect quality-damaging bit errors in the gain bits without sending redundant bits. It greatly improves the audio quality in bit error conditions without significantly affecting clear channel quality.

6. EXCITATION VECTOR QUANTIZATION

BV16 and BV32 encode the excitation using VQ with a small vector dimension of 4. The excitation VQ codebook has a simple sign-shape structure, with 1 bit for sign, and 4 bits and 5 bits for shape for BV16 and BV32, respectively. This gives equivalent codebook sizes of 32 and 64, for effective excitation encoding bit-rates of 1.25 and 1.5 bits/sample for BV16 and BV32, respectively.

The excitation VQ codebook search is performed in an analysis-by-synthesis manner, by finding the excitation VQ codevector that, when scaled and passed through the feedback filter structure in Figs. 1 and 3, gives the smallest energy of the corresponding quantization error vector $q(n)$. As shown in [5], during the excitation VQ codebook search, the feedback filter structure in Figs. 1 and 3 can be considered a linear system with the scaled VQ output $uq(n)$ as the input signal and the quantization error $q(n)$ as the output signal. With decomposition of $q(n)$ into the zero-input response (ZIR) and zero-state response (ZSR), the VQ codebook search complexity can be greatly reduced [5]. Further complexity reduction can be achieved by the techniques proposed in [6]. For an explanation of the basic concepts and detailed excitation VQ codebook search procedure, see [4], [5], and [6].

7. BIT ALLOCATION

Table 1 shows the bit allocation for BV16 and BV32 for each 5 ms frame. The side information (everything except the excitation) takes 30 bits/frame for BV16 and 40 bits/frame for BV32.

Table 1 Bit allocation of the BV16 and BV32 codecs

Parameter	BV16	BV32
LSP	7+7=14	7+(5+5)=17
Pitch period	7	8
3 pitch taps	5	5
Excitation gain(s)	4	5+5=10
Excitation vectors	(1+4)×10=50	(1+5)×20=120
Total per frame	80	160

8. POSTFILTERING AND PACKET LOSS CONCEALMENT

The BV32 codec is used without a postfilter since it provides sufficiently high output audio quality and thus does not need the quality enhancement from a postfilter. Another reason is to keep the BV32 complexity low, as adding a postfilter at 16 kHz sampling can add significant complexity. For BV16, an optional postfilter can be added at the decoder. Since postfiltering is a post-processing step, many different kinds of postfilters can be used. An example postfilter is specified in the ANSI BV16 specification [2]. Other postfilters, such as those used in many other speech coding standards, can also be used with BV16.

Packet loss concealment (PLC) is also a post-processing step at the decoder. Hence, different PLC schemes can be used with BV16 and BV32 without affecting bit-stream compatibility. An example PLC is specified in the ANSI BV16 specification [2]. This PLC can be extended for use in BV32. Alternatively, other

PLC schemes, such as those used in many other speech coding standards, can also be used with BV16 and BV32.

9. COMPLEXITY

With strong emphasis in low complexity design, BV16 and BV32 have much lower complexity than comparable codecs. While 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 DSP, the BV16 codec only takes about 12 MIPS for a full-duplex channel, a factor of 2 to 3 lower in complexity than these other codecs. Similarly, while most CELP-based wideband standard codecs such as G.722.2 (AMR-WB), VMR-WB, and G.729.1 have a complexity of around 40 MIPS, the BV32 codec only takes about 17 MIPS, which is more than a factor of 2 lower.

BV16 and BV32 also require lower memory sizes than most other comparable codecs. While most CELP-based standard codecs require 2.5 to 4.6 kwords of RAM for narrowband and 5.3 to 9.1 kwords for wideband, BV16 and BV32 only require about 2 kwords and 3 kwords of RAM, respectively. 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 and BV32 require only about 13 kwords.

10. PERFORMANCE

The output speech quality of BV16 and BV32 has been evaluated by both objective measures such as ITU-T P.862 PESQ and by formal subjective listening tests. For objective measures, PESQ and wideband PESQ have been used in a very large scale evaluation involving 13 different languages including Arabic, Chinese, English, French, German, Hindi, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, and Thai from the NTT 1994 multi-lingual speech database. For each codec evaluated, the PESQ or wideband 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 or wideband PESQ scores are averaged. This process closely resembles how the Mean Opinion Score (MOS) in a subjective listening test is calculated. The results are summarized in Table 2. It can be seen that BV16 is ranked second only to the 64 kb/s G.711 μ-law, and BV32 is ranked the highest among the 10 wideband codecs listed.

Table 2 PESQ averaged over 13 languages

Narrowband Codec	PESQ	Wideband Codec	Wideband PESQ
G.711 μ-law	4.119	**BV32**	3.788
BV16	4.077	G.722.2 at 23.85 kb/s	3.625
G.729E	4.040	G.722 at 64 kb/s	3.591
GSM-EFR	4.008	G.722 at 56 kb/s	3.514
G.726 at 32 kb/s	3.938	G.711 μ-law	3.475
G.728	3.891	G.722.2 at 15.85	3.427
G.729	3.798	G.722.1 at 32 kb/s	3.383
G.723.1 at 6.3 kb/s	3.629	G.722.1 at 24 kb/s	3.318
G.729D	3.557	G.722 at 48 kb/s	3.314
G.723.1 at 5.3 kb/s	3.489	G.722.2 at 8.85	2.903

BV16 and BV32 have been evaluated in formal subjective listening tests conducted by independent test labs Dynastat and Comsat Labs, respectively. A total of 32 naïve listeners participated in each of the two MOS tests. The resulting MOS scores are summarized in Table 3. Statistical analysis showed that in these tests, BV16 was rated statistically better than toll-quality codecs G.729, G.726 at 32 kb/s, and G.728, and BV32 was rated statistically better than G.722 at 64 kb/s. This confirms that BV16 and BV32 achieve very high output speech quality. What is not shown in Table 3 due to space limitation is that with proper PLC, both BV16 and BV32 are also robust to packet loss. While most other speech codecs degrade MOS by 0.5 relative to clear channel MOS at a random packet loss rate of 2 to 3%, formal subjective listening tests showed that BV16 and BV32 did not degrade MOS by 0.5 until the packet loss rate reached about 5%.

Table 3 MOS scores from two listening tests

Narrowband Codec	MOS	Wideband Codec	MOS
G.711 μ-law	3.91	**BV32**	4.11
BV16	3.76	G.722 at 64 kb/s	3.96
G.729	3.56	G.722 at 56 kb/s	3.88
G.726 at 32 kb/s	3.56	G.722 at 48 kb/s	3.60
G.728	3.54		

11. CONCLUSION

This paper gives a high-level description of the BroadVoice speech coding algorithm that has been standardized by CableLabs, SCTE, and ANSI for VoIP cable telephony. The BroadVoice algorithm is based on a novel two-stage noise feedback coding paradigm with efficient vector quantization of the excitation signal. With low delay, low complexity, and high audio quality as the main design emphasis, the BroadVoice16 and BroadVoice32 codecs achieve 3 to 8 times lower algorithmic delay, 2 to 3 times lower computational complexity, and more than 1/3 lower memory footprint than most other CELP-based speech coding standards. BroadVoice achieves such low delay and low complexity while maintaining equivalent or better output speech quality.

12. ACKNOWLEDGMENTS

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

13. REFERENCES

[1] J.-H. Chen and J. Thyssen, "BroadVoice®16: A PacketCable Speech Coding Standard for Cable Telephony," *Proc. Asilomar Conf. Signals, Systems, Computers*, Asilomar, CA, October 2006.

[2] "BV16 Speech Codec Specification for Voice over IP Applications in Cable Telephony," American National Standard, ANSI/SCTE 24-21 2006.

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

[4] 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.

[5] J.-H. Chen, "Novel codec structures for noise feedback coding of speech," *Proc. IEEE ICASSP*, pp. I-681 - I-684, May 2006.

[6] 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.

[7] F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," *Proc. IEEE ICASSP*, pp. 1.10.1 - 1.10.4, March 1984.

[8] R. P. Ramachandran and P. Kabal, "Stability and performance analysis of pitch filters in speech coders," *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. 35, No. 7, pp. 937 - 946, July 1987.