The BroadVoice® Speech Coding Algorithm

Juin-Hwey (Raymond) Chen, Ph.D.

Senior Technical Director
Broadcom Corporation

March 22, 2010

Outline

1. Introduction
2. Basic Codec Structures
3. Short-Term Prediction / Noise Spectral Shaping
4. Long-Term Prediction / Noise Spectral Shaping
5. Gain Quantization
6. Excitation Vector Quantization
7. Bit Allocation
8. Postfiltering and Packet Loss Concealment
9. Complexity
10. Performance
11. Conclusion

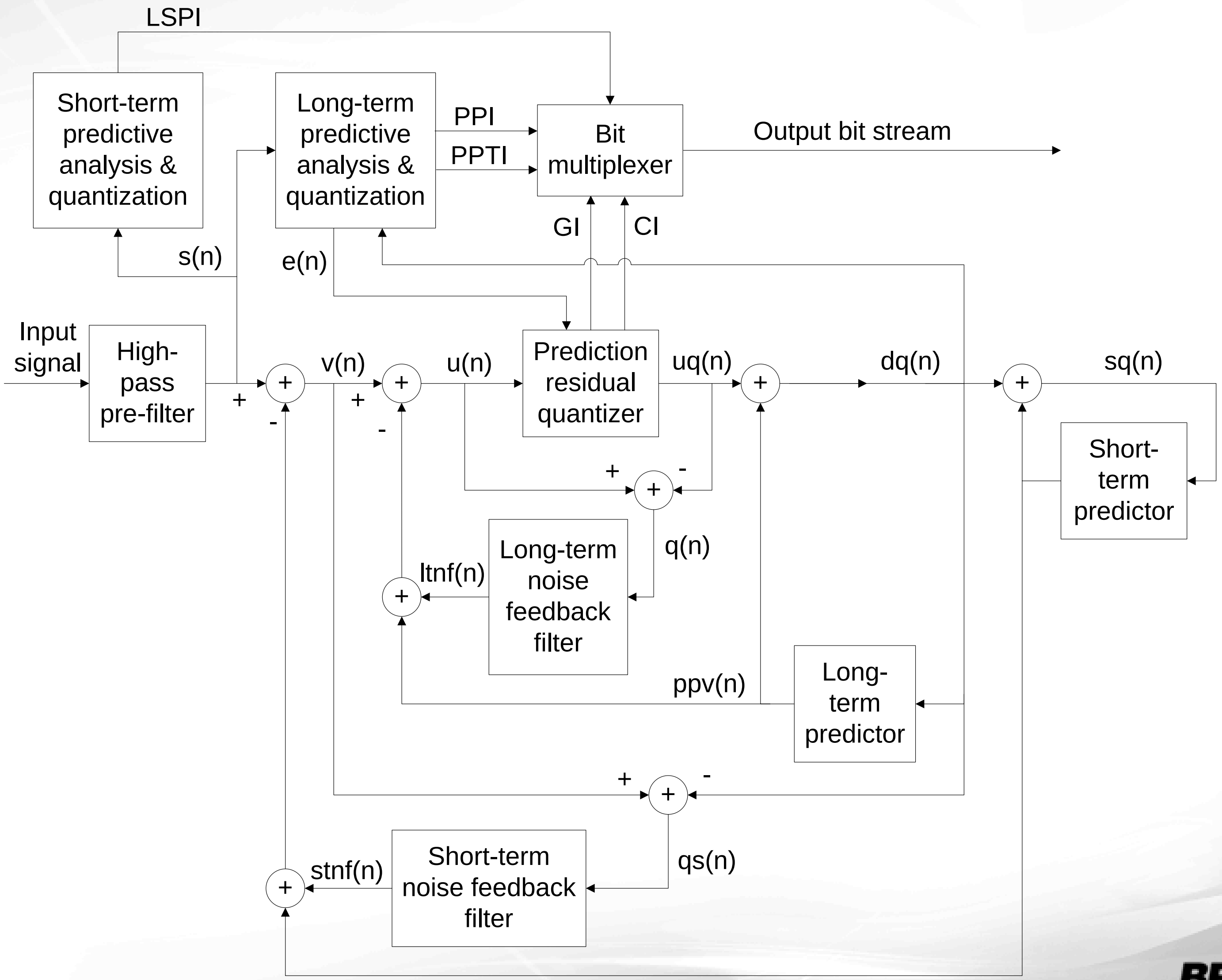
BROADCOM.
Connecting everything®

Introduction

- BroadVoice16 (BV16):
  - 16 kb/s narrowband speech codec with 8 kHz sampling
  - Selected by CableLabs in 2004 as a standard codec in PacketCable 1.5 for Voice over Cable applications; later also became a standard codec in PacketCable 2.0
  - Standardized by SCTE and ANSI in 2006 as "ANSI/SCTE 24-21 2006" standard
  - One of the standard codecs listed in the ITU-T Recommendation J.161
- BroadVoice32 (BV32):
  - 32 kb/s wideband speech codec with 16 kHz sampling
  - Standard codecs in PacketCable 2.0, "ANSI/SCTE 24-23 2007", and ITU-T Recommendation J.361
- BV16 and BV32 are:
  - based on Two-Stage Noise Feedback Coding (TSNFC)
  - optimized for **low delay**, **low complexity**, and **high speech quality**
  - Royalty-free and open source (both floating-point and fixed-point C)
  - Visit http://www.broadcom.com/broadvoice for info & code download

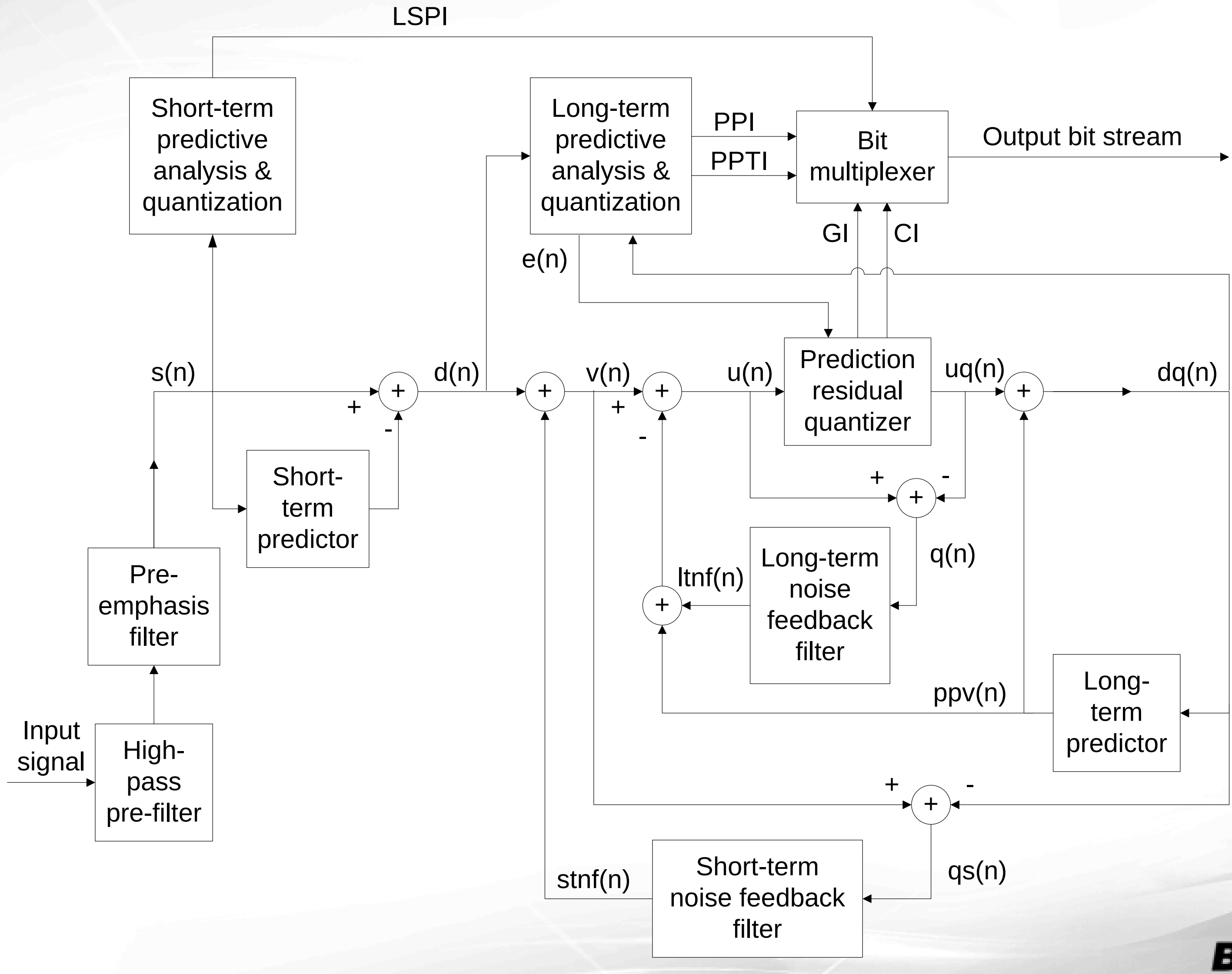
BROADCOM.
Connecting everything

BV16 Encoder Structure



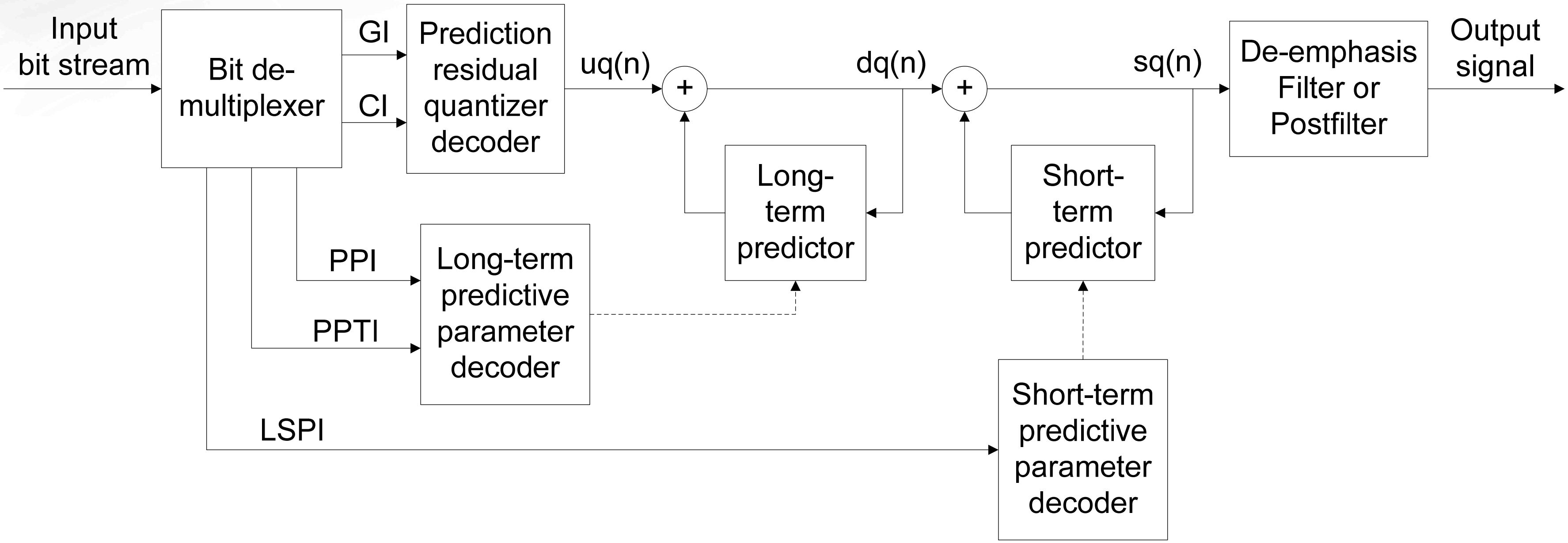
- BV16 uses TSNFC Form 3 structure in our ICASSP 2006 paper

BV32 Encoder Structure



- BV32 uses TSNFC Form 2 structure in our ICASSP 2006 paper

BV16/BV32 Decoder Structure



- Similar to a CELP decoder
- BV32 uses a de-emphasis filter but not a postfilter
- BV16 does not use a de-emphasis filter but may add a postfilter

Short-Term Prediction

- Use 8th-order short-term prediction to keep complexity low
- LSP quantized using 8th-order MA prediction and two-stage VQ:
  - 1st-stage: 8-dimensional VQ with 7-bit codebook
  - 2nd-stage: BV16 uses 8-dimensional VQ with 1-bit sign and 6-bit shape
    BV32 uses split VQ with 3-5 split and 5 bits each
- BroadVoice might be used in non-VoIP applications with bit errors
  - Desirable to make it robust to bit errors
- Only codevectors that preserve the order of first 3 LSPs are allowed in the 2nd-stage VQ codebook search
  - order reversal at decoder indicates bit errors → last LSP vector used
  - greatly reduces distortion due to bit errors without sending redundant information
  - essentially no degradation to clear-channel quality

BROADCOM.
Connecting everything®

Short-Term Noise Spectral Shaping

- TSNFC Form 2 structure of BV32 has a lower complexity but gives a more constrained noise spectral shape of

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

- TSNFC Form 3 structure of BV16 has a higher complexity but gives a more general noise spectral shape of

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

- $\tilde{A}(z)$ uses quantized coefficients while $A(z)$ uses unquantized ones

- $\gamma = 0.75$ for BV32; $\gamma_1 = 0.5$ and $\gamma_2 = 0.85$ for BV16

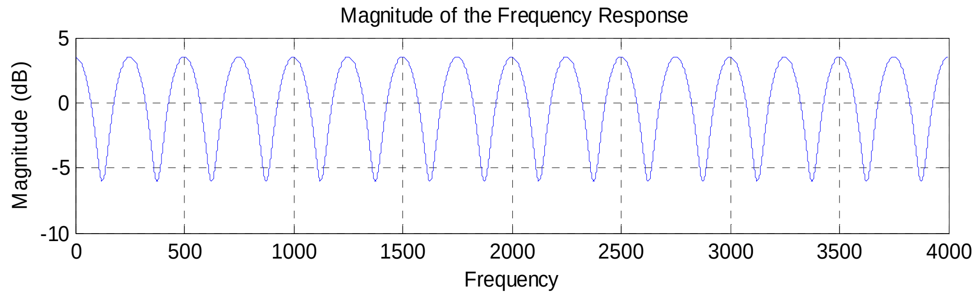
BROADCOM. Connecting everything®

Long-Term Prediction and Noise Spectral Shaping

- Long-Term Prediction:
  - 3-tap pitch predictor with integer pitch period
  - pitch period encoded to 7 bits for BV16 and 8 bits for BV32
  - pitch period range: 10 to 136 for BV16 and 10 to 264 for BV32
  - 3 pitch predictor taps vector quantized to 5 bits
  - pitch period and pitch taps determined in open-loop fashion to save complexity
- Long-Term Noise Spectral Shaping:
  - To keep the complexity low, the noise feedback filter has a simple form of

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

  - λ is half of optimal single-tap pitch predictor coefficient, range-limited to [0, 1]
  - The corresponding noise spectral shape is given by $N_l(z) = 1 + \lambda z^{-pp}$
  - Example:



Gain Quantization

- Excitation gain derived and quantized in open-loop to save complexity
- 1 gain/frame for BV16, and 2 gains/frame for BV32
- Gain: base-2 logarithm of average power of open-loop prediction residual
- Fixed moving-average (MA) prediction of gain using 40 ms worth of previous data:
  - 8^{th}-order MA predictor for BV16
  - 16^{th}-order MA predictor for BV32
- Scalar quantization of MA prediction residual of log-gain:
  - 4 bits for BV16
  - 5 bits for BV32

BROADCOM. Connecting everything®

Gain Change Limitation

- Problem: Bit errors can cause large “gain pops” in decoded speech
- Solution: Limit the maximum gain increase allowed, conditioned on the previous log-gain and previous log-gain change
  - Train a “constraint threshold matrix” off-line:
    - Row: log-gain relative to a long-term average log-gain
    - Column: log-gain change between adjacent gains
    - Matrix element values: 99.x percentile of observed log-gain change in natural speech
  - In gain encoding, if quantized gain gives a log-gain change > threshold, reduce the quantized gain until < threshold, or until the smallest gain in gain codebook
  - In gain decoding, if the gain code is not for the smallest gain in gain codebook and the decoded gain gives a log-gain change > threshold, then the gain is corrupted by bit errors → replace with the last decoded gain value
- Result: All severe “gain pops” eliminated, no redundant bit needed, and clear-channel performance hardly affected

BROADCOM. Connecting everything®

Excitation Vector Quantization

- Excitation VQ dimension = 4
  - BV16: 1-bit sign, 4-bit shape, (1+4)/4 = 1.25 bits/sample
  - BV32: 1-bit sign, 5-bit shape, (1+5)/4 = 1.5 bits/sample
  - VQ codebook closed-loop trained
- Analysis-by-synthesis codebook search:
  - concept: pass all codevectors through TSNFC structure, pick the one that gives minimum energy of quantization error
- Efficient VQ codebook search:
  - treat TSNFC structure as a linear system with VQ codevector as input and quantization error vector as output
  - decompose quantization error vector into Zero-Input Response (ZIR) and Zero-State Response (ZSR) → see our ICASSP 2006 paper
  - further complexity reduction → see our Interspeech 2006 paper

BROADCOM.
Connecting everything®

Bit Allocation

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 bits/40 samples	160 bits/80 samples

BROADCOM.
Connecting everything®

Postfiltering (PF) and Packet Loss Concealment (PLC)

- BV16 and BV32 are not bit-exact standards
- PF and PLC are both post-processing steps after decoding
- PF and PLC do not affect bit-stream compatibility
- PF and PLC are not really part of the BV16/BV32 standards
- BV16 specification gives an example PF
- BV16/BV32 specifications each gives an example PLC
- Other PF and PLC schemes can be used without affecting inter-operability with the BV16/BV32 standards

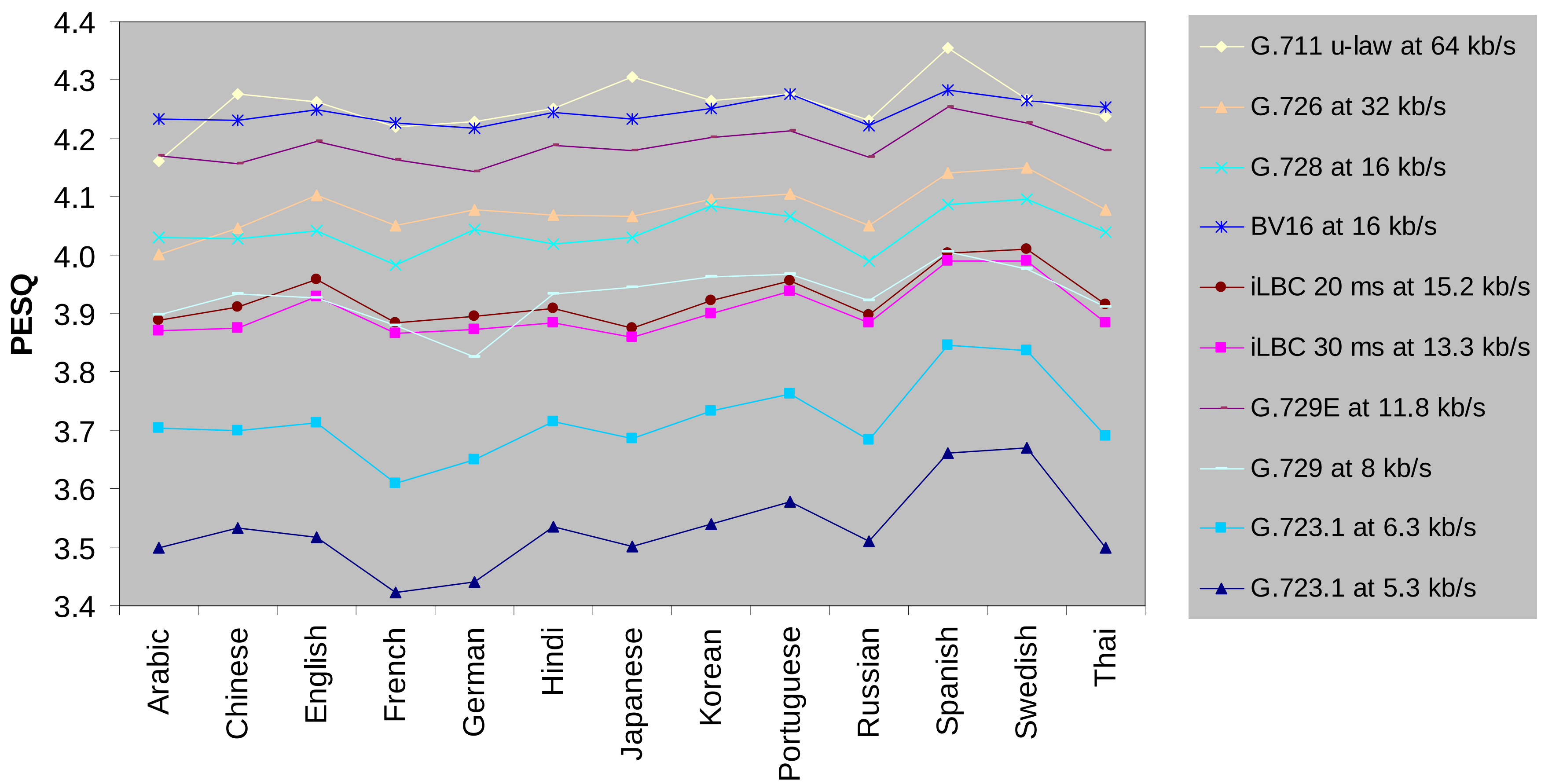
BROADCOM. Connecting everything

Complexity Comparison with Other CELP-Based Standard Codecs*

Codec	MIPS	RAM (kwords)	ROM (kwords)	Total Memory Footprint	Algorithmic Delay (ms)
G.728	36	2.2	6.7	9	0.625
G.729	22	2.6	14	17	15
G.729E	27	2.6	20	23	15
G.723.1	19	2.1	20	22	37.5
EVRC	25	2.5	?	?	30
AMR	20	4.6	17	22	25
BV16	12	2	11	13	5
G.722.2	40	5.3	18	23	26.875
VMR-WB	40	9.05	?	?	33.75
G.729.1	40	8.7	40.5	49	48.9375
BV32	17	3	10	13	5

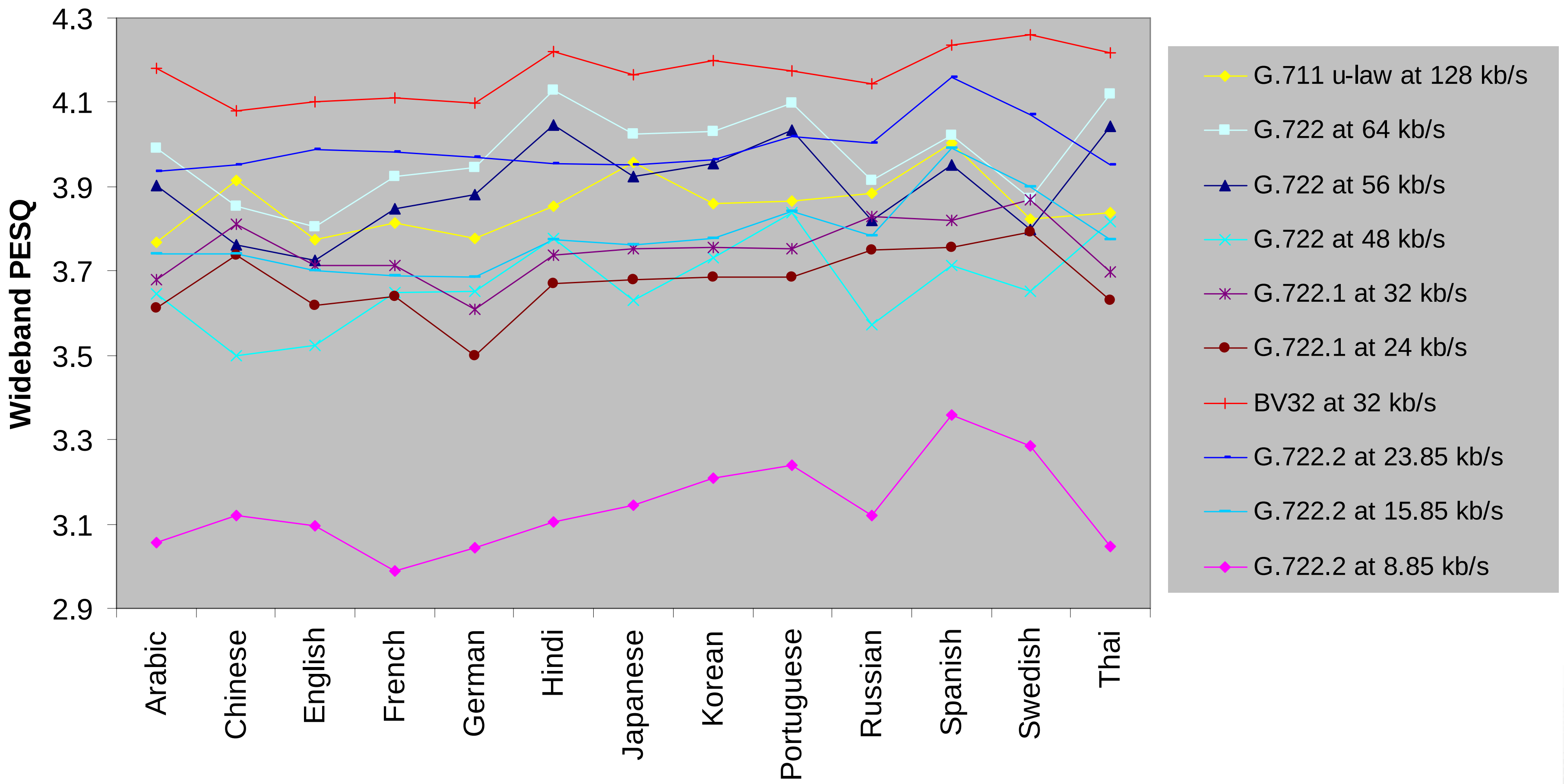
* Most data extracted from PacketCable 2.0 spec audio codec comparison table

Narrowband Speech Quality Measured by PESQ Using 13 Languages



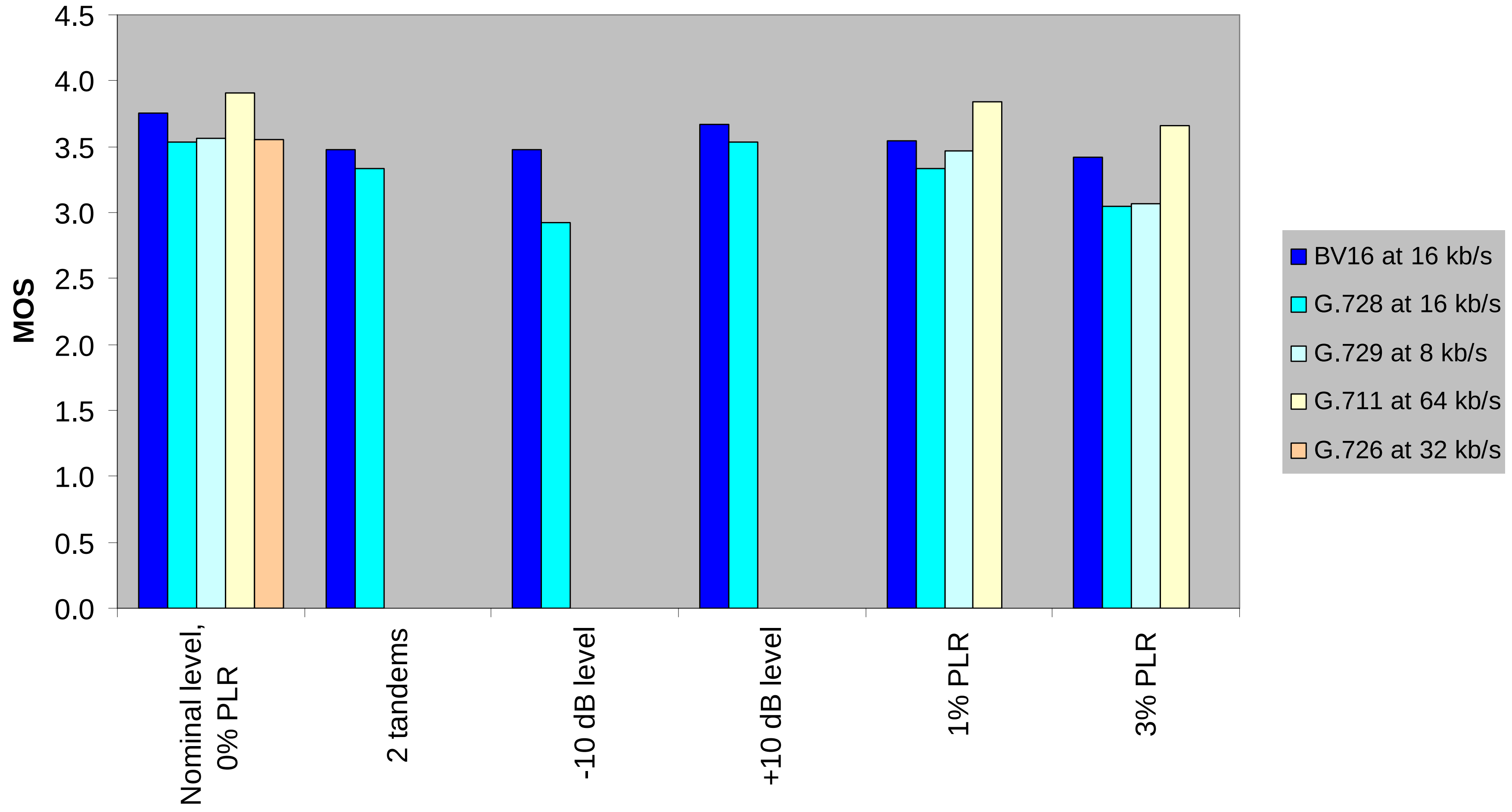
- All 96 sentence pairs of 13 languages in NTT 1994 database were used
- BV16 was rated higher than all other codecs here except 64 kb/s G.711

Wideband Speech Quality Measured by Wideband PESQ Using 13 Languages



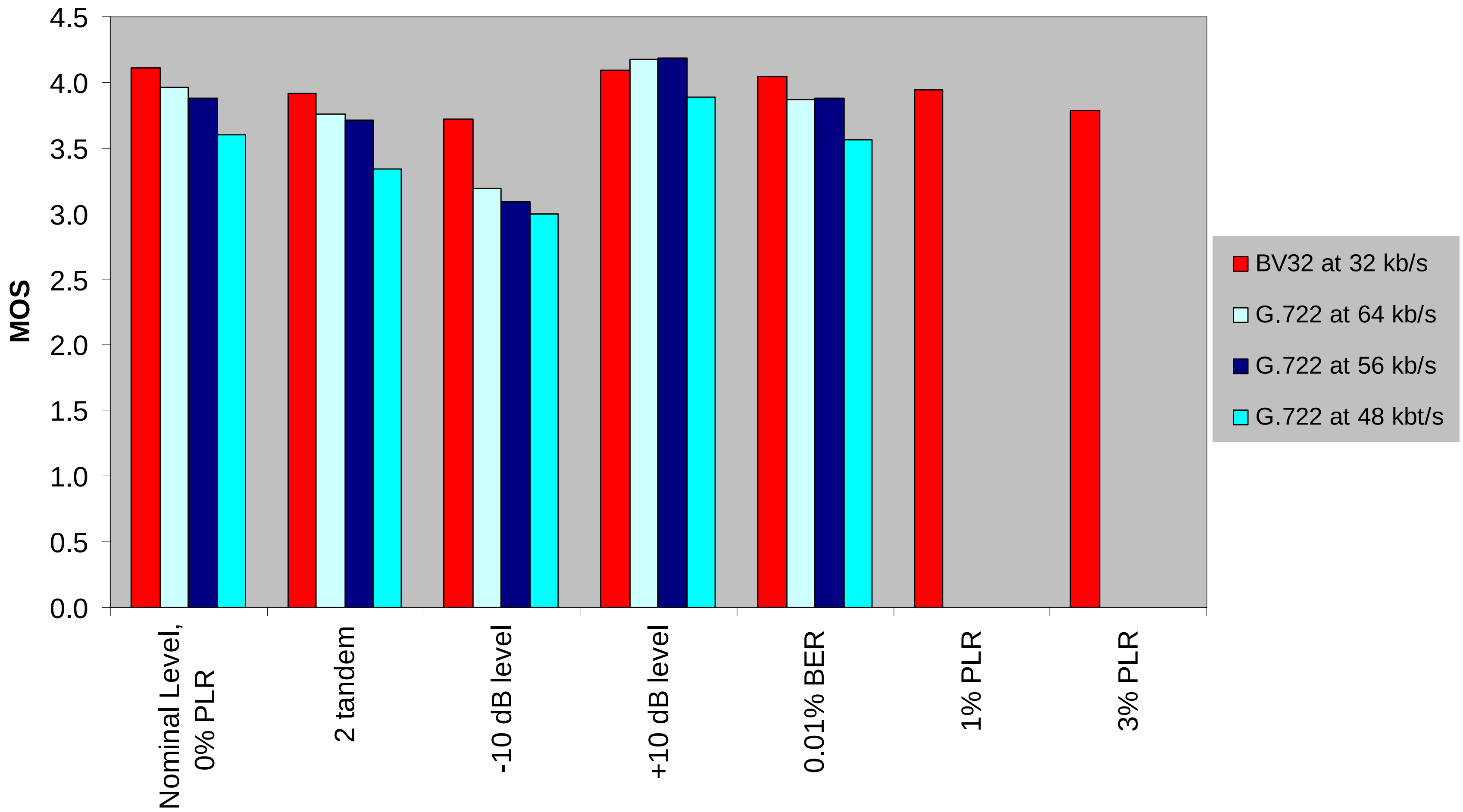
- All 96 sentence pairs of 13 languages in NTT 1994 database were used
- BV32 was rated higher than all other codecs listed here

Narrowband Listening Test Results



BROADCOM. Connecting everything

Wideband Listening Test Results



Connecting everything

BroadVoice Subjective Speech Quality Relative to Reference Codecs

- Dynastat did narrowband MOS test; Comsat Labs did wideband test
- 32 naïve listeners in each test
- BV16 rated statistically better than G.728, G.729, and G.726 at 32 kb/s
- BV32 rated statistically better than G.722 at 64 kb/s
- BV16/BV32 give 0.5 MOS degradation at about 5% random packet loss, versus 2% to 3% for most other standard speech codecs

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		

BROADCOM. Connecting everything®

Conclusion

- BroadVoice16 and BroadVoice32 are based on novel Two-Stage Noise Feedback Coding with following design emphases:
  - Low delay: 3x to 8X lower algorithmic delay than most competing codecs
  - Low complexity: 2X to 3X lower MIPS, 1.3X to 3.8X lower memory footprint
  - High speech quality:
    - BV16 statistically better than toll-quality codecs G.726 at 32 kb/s, G.728, G.729
    - BV32 statistically better than G.722 at 64 kb/s
    - Slower degradation with increasing packet loss rate than most other codecs
- BV16 and BV32 are standard speech codecs of PacketCable 1.5/2.0, ANSI, SCTE, and ITU-T J.161/J.361 for VoIP over Cable applications
- BV16 and BV32 are royalty-free and open source
- BV16 and BV32 can potentially be a base layer codec of IETF Internet Interactive Audio Codec → benefit: can make IIAC inter-operable with existing ANSI/SCTE BV16/BV32 standards

BROADCOM.
Connecting everything®