Improved Noise Weighting in CELP Coding of Speech

Applying the Vorbis Psychoacoustic Model To Speex

By: Jean-Marc Valin, Christopher Montgomery
22/5/2006

CeNTIE is supported by the Australian Government through the Advanced Networks Program (ANP) of the Department of Communications, Information Technology and the Arts and the CSIRO ICT Centre

CSIRO

Introduction

- Goal: Improve perceptual weighting of the noise in an existing CELP codec (Speex)
- Proposed solution: adapt and apply the Vorbis psychoacoustic model to the Speex codec
- Outline
  - Overview of Speex
  - Overview of Vorbis and psychoacoustic model
  - Application to Speex
  - Evaluation & results
  - Complexity
  - Conclusion

Overview of Speex

- Speech codec based on CELP
- Sampling rates, bitrates:
  - Narrowband (8 kHz): 2.15 kbps to 24.6 kbps
  - Wideband (16 kHz): 3.95 kbps to 42.2 kbps
- Features:
  - Open-source (BSD-licensed): http://www.speex.org/
  - Source-controlled variable bitrate (VBR)
  - Embedded wideband coding
  - Variable encoder complexity
  - Optimised for VoIP
- Bit-stream finalized in March 2003

Speex Encoder Structure

- CELP variant with
  - 20 ms frames (5 ms sub-frames)
  - No inter-frame coding other than LPC and pitch prediction
  - 3-tap pitch predictor
  - Sub-vector quantization of innovation
  - “Global” excitation gain
- Default noise weighting is LPC-derived

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \gamma_1 = 0.9, \gamma_2 = 0.6$$

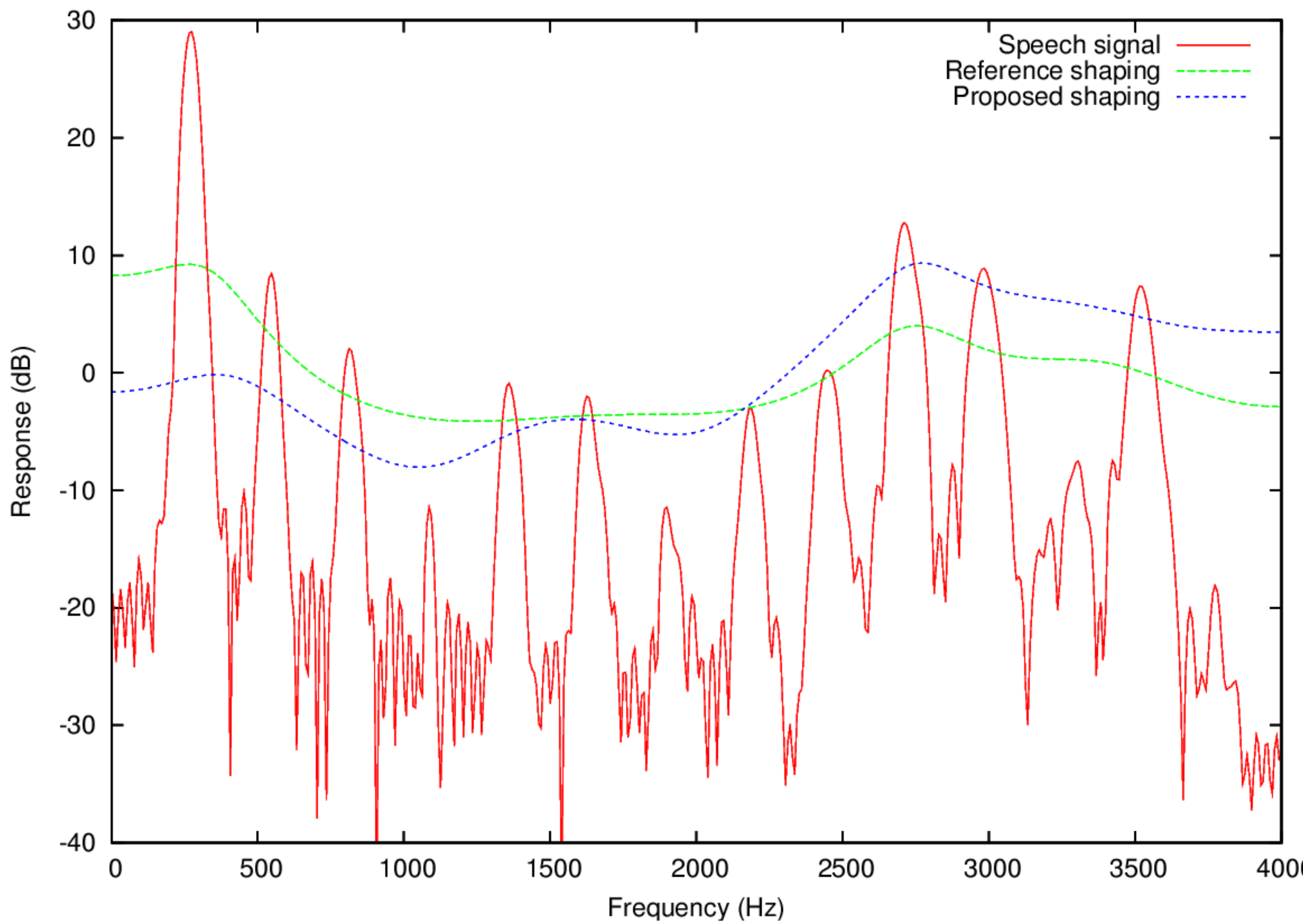
Vorbis Psychoacoustic Model

- Vorbis is an open-source, MDCT-based audio codec
- Psychoacoustic model shapes noise according to:
  - Tone masking
  - Noise masking
  - Noise normalization
  - Impulse analysis
- Noise shaping approximates the masking threshold
  - Good for transparent audio
  - Bad for lossy speech

Application to Noise Weighting in Speex

- Vorbis "floor" curve interpreted as the inverse of the optimal perceptual weighting filter
  - Amplitude companding required
- Compute curve for each frame and interpolate on sub-frames
- Convert to pole-zero model: $\frac{1}{W(z)} = \frac{W_n(z)}{W_d(z)}$
  - Denominator:
    - Curve to auto-correlation (IFFT)
    - Auto-correlation to LPC (Levinson-Durbin)
  - Numerator:
    - Remove denominator contribution (1/FFT of denominator)
    - Convert inverse to LPC (IFFT and Levinson-Durbin)

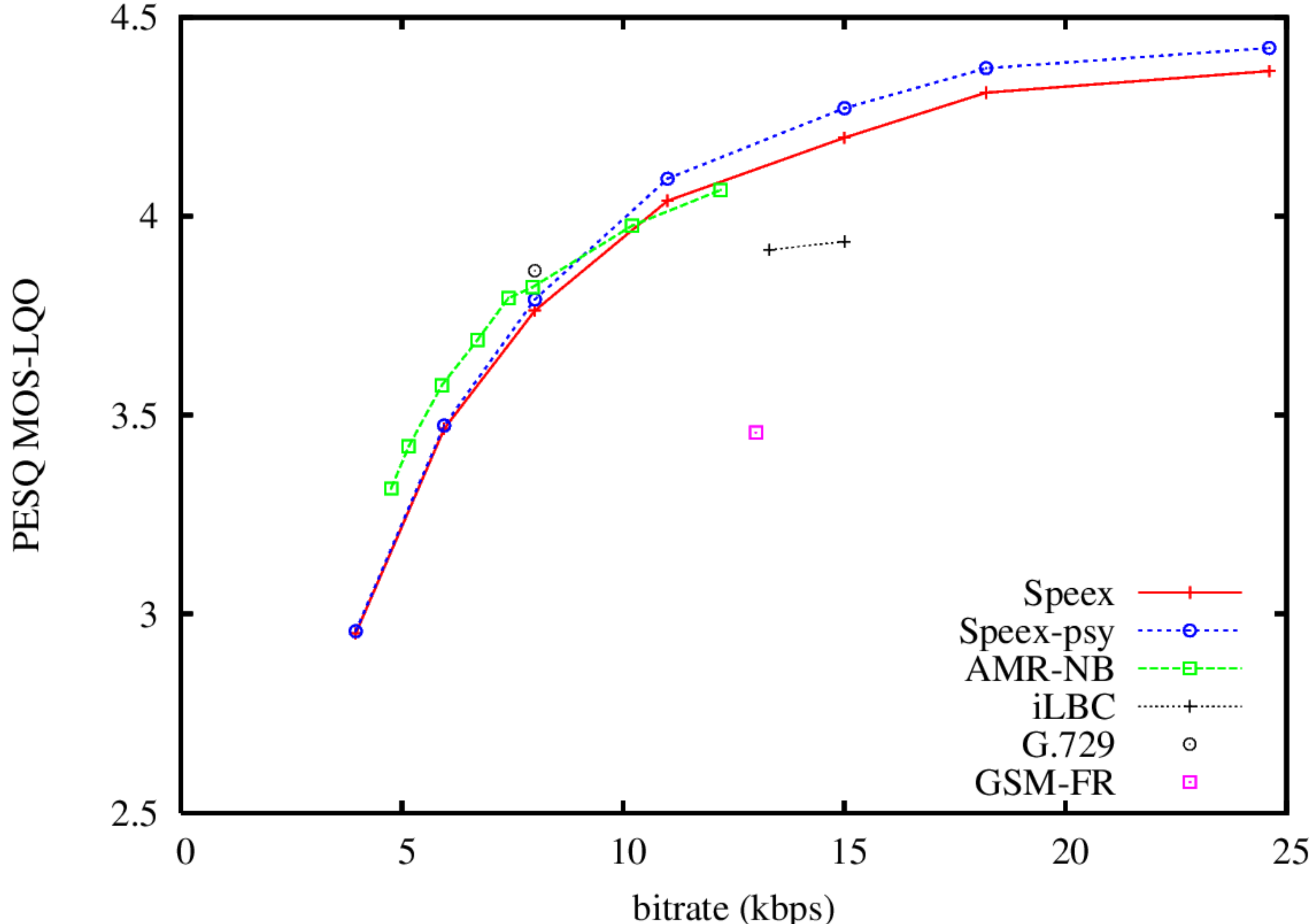
Curves


Speech signal
Reference shaping
Proposed shaping
Response (dB)
Frequency (Hz)
30
20
10
0
-10
-20
-30
-40
0
500
1000
1500
2000
2500
3000
3500
4000

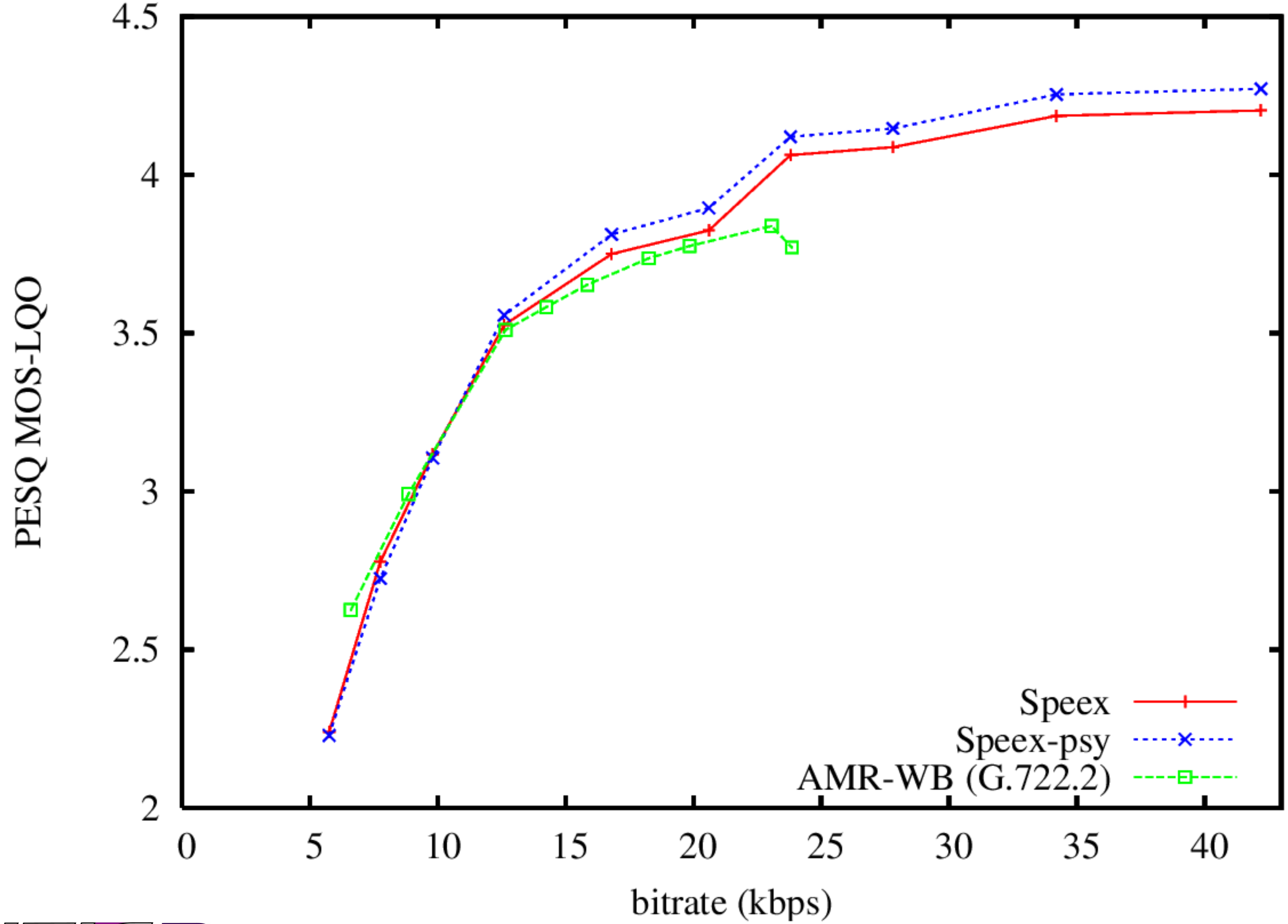
Evaluation

- Objective listening quality: PESQ MOS-LQ0 (P.862.x)
- Tested on NTT multilingual speech database
  - 354 files
  - 177 speakers
  - 20 languages
- Reference: Speex version 1.2-beta1 (pre-release)

Results (narrowband)



Results (wideband)



CeNTIE2

CSIRO

Complexity Reduction

Three strategies:

1) Use all-pole model $\frac{1}{W(z)} = \frac{1}{W_d(z)}$

2) Force $W_d(z) = A(z)$

- Synthesis+weighting filter simplifies to $\frac{W(z)}{A(z)} = \frac{1}{W_n(z)}$
- Reduces complexity of the filtering

3) Apply 2) and make $W_n(z)$ constant for a whole frame

- Only one conversion per frame

None of 1), 2) or 3) causes significant degradation

Conclusion

- Proposed an improved noise weighting for the Speex codec
- Noise weighting is based on the Vorbis psychoacoustic model
- Up to 20% (equivalent) improvement at high bitrate
- Little or no improvement at low bitrate
- **A case for more research to be done in noise weighting for CELP**
- A subjective MOS test is desirable
- Future work
  - Investigate efficient approximations for $W_n(z)$
  - Derive CELP-specific masking models

Questions?