Perceptual Coding of Narrowband Audio Signals at 8 kb/s

Hossein Najafzadeh   *Peter Kabal*

Telecommunications & Signal Processing Laboratory

McGill University

Motivations & Objectives

- Rapid growth of multimedia communications (wireless PCS & internet) requires efficient algorithms for coding and reproduction of audio signals
- Efficient use of the bandwidth requires low rate coding
- Use the *masking property* of the hearing system in the coding algorithm; the distortion introduced in the coding process is masked by the input signal
- **Goal**: develop a coding algorithm for narrowband audio inputs using 1 bit/sample with acceptable quality

Proposed Coder Overview

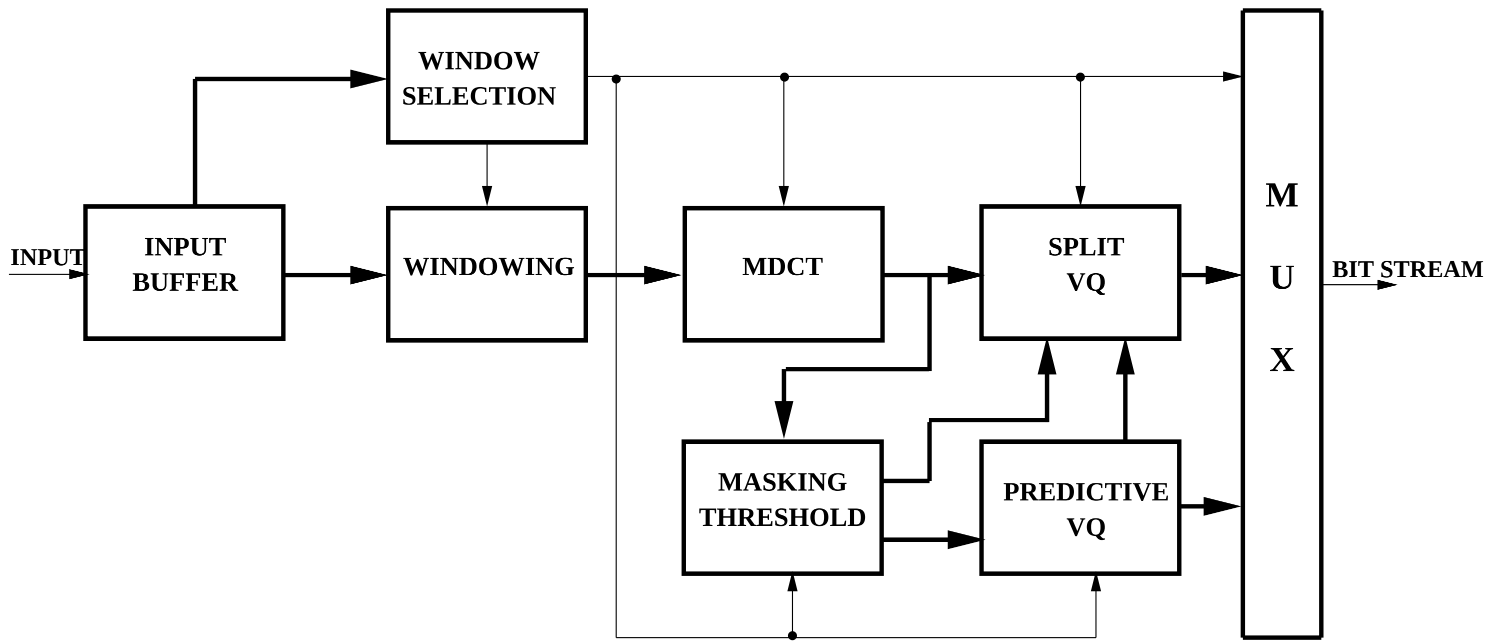

Fig. 1 Block diagram of the proposed coder.

- *Adaptive Time to Frequency Mapping*
  - use an MDCT with 50% overlap between successive frames
  - for blocks with no sharp transients, a frame of 240 samples (30 msec) is transformed into 120 coefficients
  - to reduce *pre-echo* artifacts a shorter window with a length of 10 msec is used whenever a strong transient is detected
  - a start window is used to switch from long to short windows, and a stop window switches back

- *Window Selection*
  - window selection is done based on the energy-entropy criterion proposed by Sinha and Tewfik as follows
  - each block of 240 samples is divided into 80 segments of 3 samples
  - energy-entropy defined as:

$$I = -\sum_{i=1}^{80} \sigma_i^2 \log_2 \sigma_i^2$$

  σ_i^2 is the energy of segment i normalized by the overall frame energy
  - a value of $I < 2.5$ bits is used as the threshold for switching

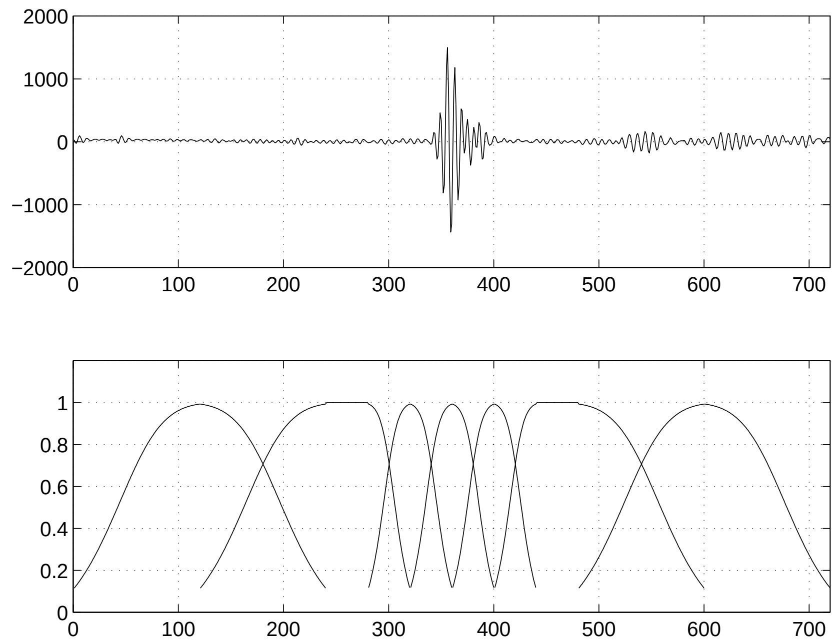

Fig. 2 Window switching for a piece of music containing a sharp jump.

- *Masking Threshold*: calculated based on the model proposed by Johnston:

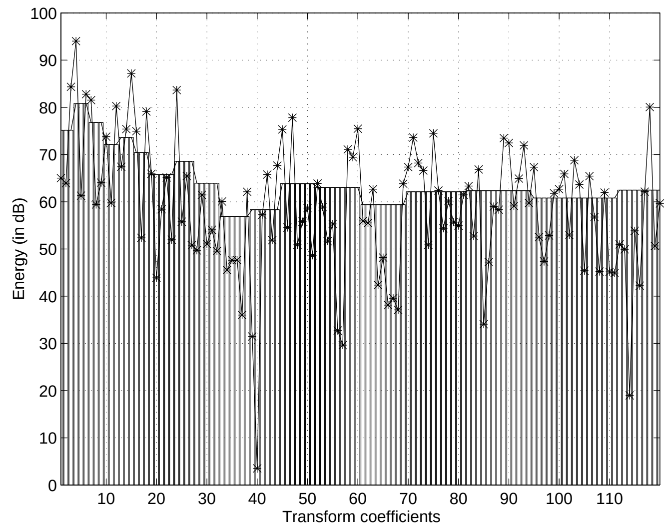

Fig. 3 Transform coefficients and masking threshold for a frame of music.

- *Perceptually Trained VQ*
  - training of the codebooks is done using a perceptual distortion measure based on the energy of the unmasked noise

$$e(i) \stackrel{\triangle}{=} |X(i) - C^{(j)}(i)|^2 - M(i)$$

$$D(X, C^{(j)}) = \sum_{i=1}^{N} (|e(i)| + e(i))/2$$

$$C_{\mathrm{opt}}^{(j)} = \arg\min_{C^{(j)}} \sum_{k=1}^{L} D(X^{(k)}, C^{(j)})$$

  - use the same criterion to search for the best codeword

- *Bit Assignment Scheme*
  - bit assignment is performed based on the distribution of the energy above the masking threshold
  - for each band an experimental relation between the distortion and the quantized energy for different bits has been found as:

$$D_i = c\,\hat{E}\,2^{-b_i/\alpha}, \quad 0.4 \le c \le 0.7, \quad 0.3 \le \alpha \le 0.5$$

  - number of bits assigned to each band is determined by:

$$\arg\min_{b_i} \sum_{i=1}^{18} D_i, \quad s.t. \quad \sum_{i=1}^{18} b_i = B$$

B is the total number of bits for each frame.

$$b_i = 5 + \alpha_i \log_2(\hat{E}_i/\hat{E}_{gm})$$

$$\hat{E}_{gm} = \left(\prod_{i=1}^{18} \hat{E}_i^{\alpha_i}\right)^{\left(1/\sum_{i=1}^{18} \alpha_i\right)}$$

- *Predictive VQ of E's*
  - energy vectors E's are highly correlated $\Rightarrow$ Predictive VQ
  - E is normalized to a unit energy vector E_n
  - an estimate of the current normalized vector is obtained using the 6 previous normalized vectors:

$$\arg\min_{c_i \in \mathcal{C}} \sum_{j=1}^{6} (E_n^{(j)} - \sum_{i=1}^{6} c_i \hat{E}_n^{(j-i)})^2$$

$$\tilde{E_n}^{(j)} = \sum_{i=1}^{6} c_i \hat{E}_n^{(j-i)}$$

$\mathcal{C}$ is the predictor codebook

$$r^{(j)} = E_n^{(j)} - \tilde{E_n}^{(j)}$$

$r^{(j)}$ is quantized using a 2 stage VQ.

$$\hat{E_n}^{(j)} = \tilde{E_n}^{(j)} + \hat{r}^{(j)}$$

Bit Allocation Table

window flag	1 bit
predictor coefficients	9 bits
residual	22 bits
EAM norm	5 bits
transform coefficients	83 bits
Total	120 bits

window flag	1 bit
normalized EAM	10 bits
EAM norm	5 bits
transform coefficients	24 bits
Total	40 bits

Table 1 Bit allocation for long and short frames

Results

- *Subjective testing*: proposed coder, compared with two low rate coders (ITU–T G.729 and EIA/TIA IS–96 at 8 kbit/sec), gives a better quality for most audio inputs, except single speaker (all coders have comparable performance)
- *Objective testing*: define a perceptually based objective measure, Signal to Audible Noise Ratio (SANR).

$$\mathrm{SANR} = \frac{\sum_{j=1}^{L} \| X^{(j)} \|^2}{\sum_{j=1}^{L} D^{(j)}}$$

File	Coder	SNR (dB)	SEGSNR (dB)	SANR (dB)	Subjective rank
Female Vocal	Proposed	13.70	13.73	20.75	1
	EIA/IS–96	11.10	11.62	13.60	2
	ITU/G.729	6.54	6.79	6.62	3
Symphony Orchestra	Proposed	9.11	9.12	14.66	1
	EIA/IS–96	6.44	6.59	8.50	2
	ITU/G.729	0.18	0.77	0.66	3
Female Speech	Proposed	10.35	9.49	14.38	≈ 3
	EIA/IS–96	7.92	6.86	9.29	≈ 2
	ITU/G.729	5.24	2.99	5.79	≈ 1

Table 2 Objective and subjective comparison for different coders

Conclusions

- We have developed a transform audio coder suited for a wide range of inputs at 8 kbit/s.
- The proposed coder delivers acceptable quality for most audio signals while other state-of-the-art speech coders operating at the same rate have uneven results for non-speech signals.
- This work has revealed the suitability of VQ–based perceptual coding systems at a rate of 8 kb/s.