Aalto University

Postfiltering Using Log-Magnitude Spectrum for Speech and Audio Coding

Sneha Das[1] **and Tom Bäckström**[1]

[1] *Department of Signal Processing and Acoustics, Aalto University, Finland*

September, 2018

Introduction

- Performance of advanced frequency domain codecs deteriorate at low-bitrates.
  - Fewer bits for encoding, thus regions with lower energy tend to be quantized to zero.
  - Speech signals have lower energy at higher frequencies.
  - Large parts of speech quantized to zero.
  - Yields spectral holes, rendering a perceptually distorted and muffled characteristic to the signal.

Current Solutions

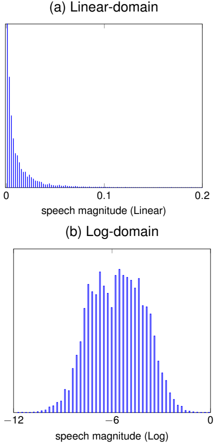
- Pre- and Post-filtering methods used to mitigate this problem. Examples are:
  - Formant enhancement
  - Bass postfilter
  - Noise filling
- Some methods are applied at decoder only and does not require any changes to the core structure of the codec, while others need to be implemented both at the encoder and decoder.
- Transmission of additional side information.
- Current solutions focus mainly on solving the manifestation of problem.

Features of proposed method

- Post-filtering method, thus applied only at the decoder, without the need for any changes to the codec structure.
- Incorporates inherent speech correlations to estimate the lost information (focusing on the cause of the problem).
- Operates using (i) quantized signal as the noisy observation, (ii) statistical models trained offline.
- Transmission of additional side information *not* required.

Speech Magnitude-Spectrum Models
Distribution

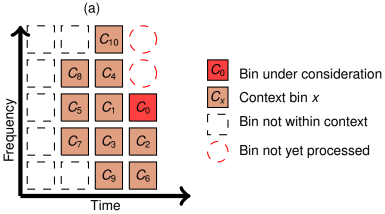
- Spectral magnitude envelope contains formant information.
- Magnitude Spectrum of speech is an exponential distribution, concentrated at low values.
    - Modeling gives rise to numerical inaccuracies.
    - Difficult to ensure positivity of estimates, using generic mathematical operations.
- *Log-magnitude spectrum:* redistribution of magnitude axis (non-linear operation of logarithm)$\rightarrow$ approx. Gaussian distribution.



Speech Magnitude-Spectrum Models

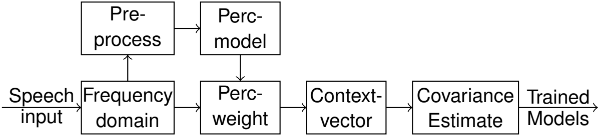
Context neighborhood

- Speech is a slowly varying signal→temporal correlation.
- Context neighborhood: surrounding frequency bins



- C_0 currently under consideration.
- Context size $= 10$ is depicted.
- Bins chosen based on distance from bin under consideration.
- Only previously estimated bin information included in context.

Speech Magnitude-Spectrum Models
Training Overview



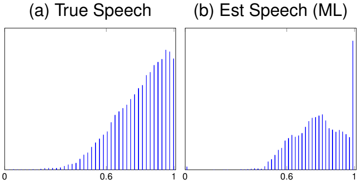
- Time-domain signal transformed to frequency domain.
- Pre-processing step includes whitening.
- Signal transformed to the perceptual domain using perceptual weighting, in accordance to CELP.
- Context-vectors of the desired size is extracted and covariance computed.

Problem Formulation I

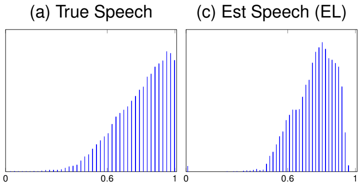
- Maximizing the likelihood of current sample, given the noisy observation and the previous estimates.

$$\hat{x} = \arg\max_{x} \quad P(X|\vec{X}_c = \hat{\vec{x}}_c) \quad \text{subject to}_{l \leq X \leq u}, \quad Q \in [l, u]$$

- Edge-problem: the estimates are biased towards the limits of the quantization-bin.



Problem Formulation II



- *Expected likelihood:*

$$\hat{x} = \arg \max_{x} E[P(X|\vec{X}_c = \hat{\vec{x}}_c)] \quad \text{subject to.} \; l \leq X \leq u \tag{1}$$

- *Truncated Gaussian model* employed for analytical solution.

Algorithm

Algorithm 1 Estimation of signal from quantized observation

Require: Quantized signal Y, speech-models C
1: **function** ESTIMATION(Y, C)
2:     **for** $frame = 1 : N$ **do**
3:         **for** $bin = 1 : Length(Y(frame))$ **do**
4:             $\mu_{up}, \sigma_{up} \leftarrow UpdateStatistics(C, \hat{X}_{prev})$
5:             $pdf \leftarrow TruncateGaussian(\mu_{up}, \sigma_{up}, l(bin), u(bin))$
6:             $\hat{X} \leftarrow Expectation(pdf)$

Systems Overview

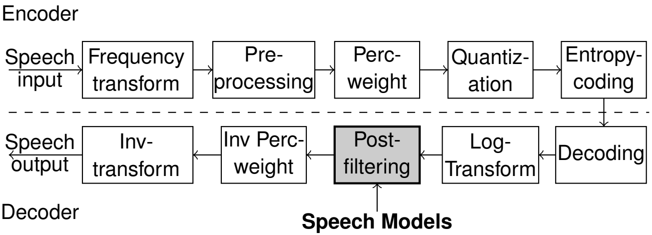
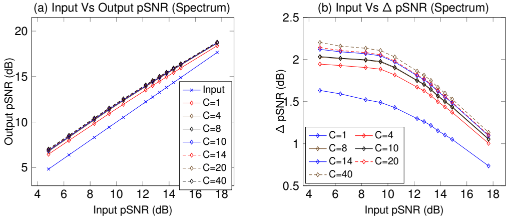


Figure: Systems block diagram

Experimental Setup

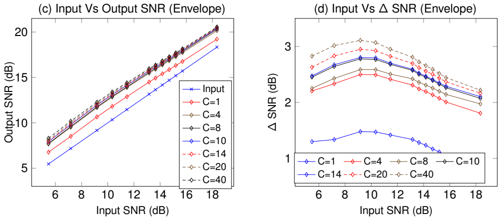
- Effect of contextual statistical models investigated in terms (a) magnitude spectrum (b) spectral envelope:
  - Cepstral coefficients used to model and estimate spectral envelope.
- Experimental setup: TIMIT database used for training and testing.
  - Training: 250 speech samples randomly chosen from the training set.
  - Testing: 10 speech samples randomly chosen from the test set, coded at 12 different bitrates between 9.6-128 kbps
  - Postfilter with context size $\in \{1, 4, 8, 10, 14, 20, 40\}$ applied to each test case.

Results
Plots I: Magnitude spectrum



- Low input pSNR: improvements in range 1.5-2.2 dB.
- Higher input pSNR: improvements in range 0.2-1.2 dB.
- Improvement in quality is large between context of size 1 and 4.

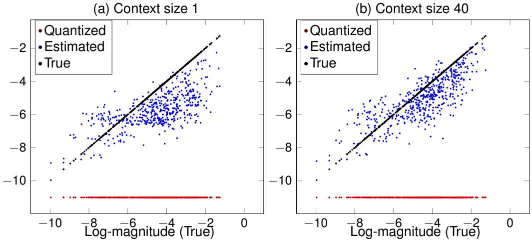
Results
Plots II: Spectral Envelope



- Low input SNR: improvements in range 1.25-2.75 dB.
- Higher input SNR: improvements in range 0.5-2.25 dB.
- Similar trend between spectrum and envelope indicate that Gaussian distributions pre-dominantly incorporate spectral envelope information.

Results

Plots III: Correlation between true and estimated speech



- Scatter plots represent the correlation between true, estimated and quantized speech in bins quantized to *zero*.
- Correlation between estimated and true values improve with increase in context size.

Conclusions

- We investigated the use of contextual information inherent in speech for the reduction of Quantization-noise.
- The proposed method employs statistical models at the decoder to estimate spectral magnitude, without the transmission of any additional information.
- Results demonstrate an average 1.5 dB improvement for inputs in the range of 4 – 18 dB and improvement prominent in bins quantized to zero.
- Besides improving pSNR for coding, the method provides spectral magnitude estimates for noise filling algorithms.