Aalto University

Postfiltering with Complex Spectral Correlations for Speech and Audio Coding

Sneha Das[1] and Tom Bäckström[1]

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

September, 2018

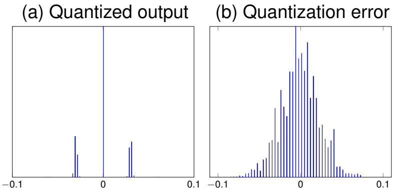
Introduction

- Speech signals $\rightarrow$ dominated by low-energy components (high frequencies).
- Encoding at low bitrates: Sparse signal (low-energy parts quantized to zero).
- Signal distorted, noise referred to as musical noise.
- Pre- and post-processing methods employed to mitigate this problem.
  - some need to be implemented both at encoder and decoder, thus modifying core codec structure.
  - some methods need to transmit additional side information.

Proposed method

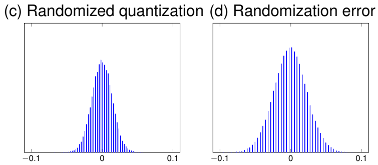
- Speech is slowly varying = high temporal correlation.
- Speech temporal and frequency correlation show noise-reduction potential.
- Speech codecs avoid transmitting information with temporal dependency; so not sufficiently studied.
- In this work, we propose:
  - a postfilter using speech models, applied at the decoder only to reduce quantization noise.
  - Models incorporate the complex spectrum characteristics.
  - Postfilter optimal in MMSE sense.

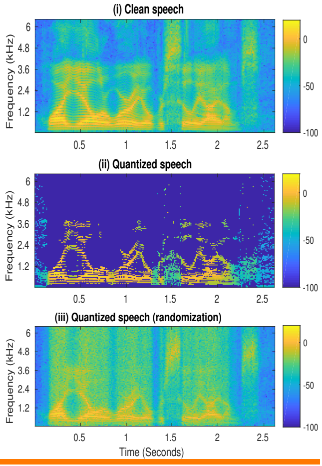
Quantized signal and Quantization noise
Characteristics



- Quantized signal is sparse $\implies$ distribution shifts away from true signal distribution.
- Quantization noise highly correlated to the original signal.

Quantized signal and Quantization noise
Dithering

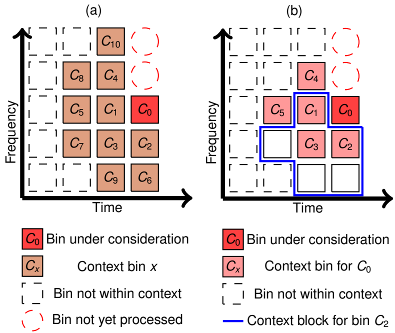




- Randomization: type of dithering.
- Dithering preserves the quantized signal distribution.
- Also lends the quantization noise more uncorrelated characteristic.

Speech Modeling
Context Neighborhood

- Context: surrounding frequency bins.
- (a) Context neighborhood of size $L = 10$.
- (b) Recursive integration of context information , similar to IIR filtering.



Speech Modeling
Problem Formulation

- Signal model: $Y_{k,t} = X_{k,t} + V_{k,t}$.
- We assume speech, X, and noise, V, are zero-mean Gaussian random variables.
- We maximize the likelihood of the clean speech estimate, $\hat{\mathbf{x}}$, given the observation $Y_{k,t}$ and the context $\hat{\mathbf{x}}_L$, such that $Y_{k,t} = \hat{X}_{k,t} + \hat{V}_{k,t}$ (constraint)
- Thus, the Optimal Wiener filter

$$\hat{\mathbf{x}} = \Lambda_X(\Lambda_X + \Lambda_N)^{-1}\mathbf{y}, \tag{1}$$

$\Lambda_X, \Lambda_N \in \mathbb{C}^{(L+1)\times(L+1)}$, L: context length.

Speech Modeling
Normalized covariance and gain modeling

- Speech signals undergo large fluctuations in gain and spectral envelope structure.
- We remove the effect of this gain using normalization during offline modeling, to obtain the static speech covariance models, Λ_X.
- The gain, γ, is computed during noise attenuation.
- Thus, the estimate of the current sample is obtained employing both Λ_X and γ, $\hat{\Lambda}_X = \gamma\Lambda_X$, $\hat{\Lambda}_X$ is the dynamic covariance model.

Systems Overview

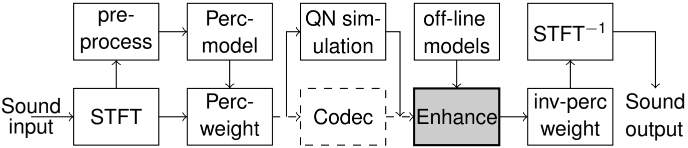


Figure: Block diagram of the proposed system.

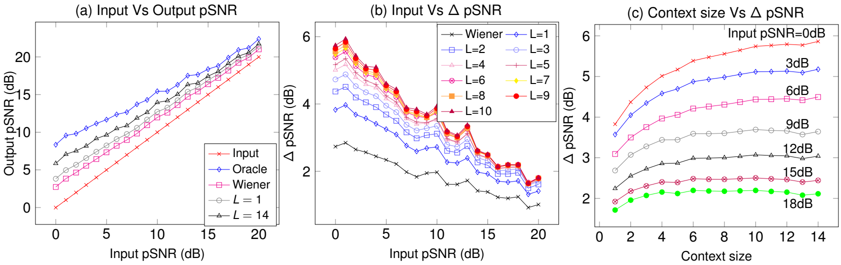
Results
Objective evaluation I

Experimental setup

- Training:
  - 50 speech signals randomly chosen from TIMIT test dataset.
  - We resample the signals to 12.8 kHz and apply Sine window on frames of 20 ms with 50% overlap and transform to the frequency domain.
  - Modeling applied in the perceptual domain.
- Testing:
  - 105 speech samples randomly chosen and noisy signals generated by adding perceptually weighted noise to obtain signals in pSNR range 0-20 dB, 5 samples for each pSNR level.
  - We tested postfilters using context sizes from 1-14.
  - Ideally enhanced signal (known noise energy) was used as reference.

Results
Objective evaluation II

Evaluation results:



Results
Subjective evaluation I

Experimental setup

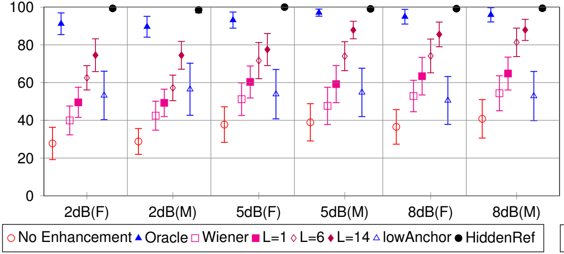
- MUSHRA test:
  - Test comprised of 6 items and 8 test conditions.
  - Experts and non-experts in the age group 20-43 were included in the test.
  - 15 listeners (9 out of total 24 listeners were discarded after failure to identify hidden reference).

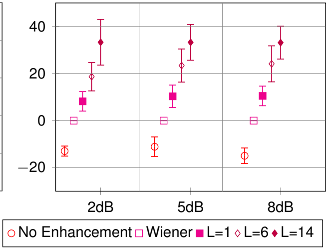
Results II
Subjective evaluation I

- Test set:
  - 6 random sentences from the TIMIT test dataset.
  - Noisy sentences with additive perceptual noise at SNR=2, 5 and 8 dB..
  - Male and female noisy cases were tested for each pSNR.
  - Conditions: hidden reference, lower-anchor, noisy, ideal enhancement, conventional Wiener filter, proposed postfiltering with context sizes 1, 6, 14.

Results
Subjective evaluation II

Evaluation results:





Conclusions

- Presented a time-frequency filter for attenuation of quantization noise.
- The complex speech correlations are modeled offline and used at the decoder only, thus eliminating the chances of error propagation from transmission loss.
- Objective tests indicate an improvement of 6 dB in best-case scenario, and 2 dB in a typical application.
- Subjective results show an improvement of 10-30 MUSHRA points.