Aalto University
School of Electrical
Engineering

INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI, CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

Fundamental Frequency Model for Postfiltering at Low Bitrates in a Transform-Domain Speech and Audio Codec

Sneha Das[1], **Tom Bäckström**[1], **Guillaume Fuchs**[2]

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

[2] *Fraunhofer IIS, Germany*

20.10.2020

Motivation

- Speech coding $\rightarrow$ Enables speech transmission $\rightarrow$ Optimize resource consumption for transmission + transmitted speech quality.
- Postfilters $\rightarrow$ Improve signal quality at decoder.
- Conventional postfilters $\rightarrow$ (a) Processing at both encoder and decoder, (b) Additional transmitted bits, (c) Dependent on other codec functional blocks.
- Alternate approach $\rightarrow$ Include source models at the decoder.
- This work $\rightarrow$ Low-complexity postfilter operating at the decoder side, which includes speech fundamental frequency models.

METHODOLOGY

Signal Model

- Speech $\rightarrow$ Glottal excitation filtered by Vocal tract response.
- Filtering in linear domain $\rightarrow$ Multiplicative $\implies$ Additive in log-domain $\rightarrow$ $\log|\mathbf{s}| = \mathbf{x}_{\mathbf{F_0}} + \mathbf{x}_{\mathbf{env}}$, $\mathbf{x}_{\mathbf{F_0}} \rightarrow$ excitation, $\mathbf{x}_{\mathbf{env}} \rightarrow$ spectral envelope.
- We model the decoded signal as $\mathbf{y}$ as follows: $\log|\mathbf{y}| = \mathbf{x}_{\mathbf{F_0}} + \mathbf{x}_{\mathbf{env}} + \mathbf{x}_{\mathbf{n}}$.
- Goal: find $\mathbf{A} = [\mathbf{A}_{\mathbf{F_0}}, \mathbf{A}_{\mathbf{env}}, \mathbf{A}_{\mathbf{n}}, \mathbf{b}]^T \rightarrow \hat{\mathbf{s}} = \mathbf{A}_{\mathbf{F_0}}\mathbf{x}_{\mathbf{F_0}} + \mathbf{A}_{\mathbf{env}}\mathbf{x}_{\mathbf{env}} + \mathbf{A}_{\mathbf{n}}\mathbf{x}_{\mathbf{n}}$.
- Optimization: minimize mean square error $\rightarrow \mathbf{A} = (\mathbf{D}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{S}^T$, $\mathbf{D} \rightarrow$ feature matrix.

System Overview

Features

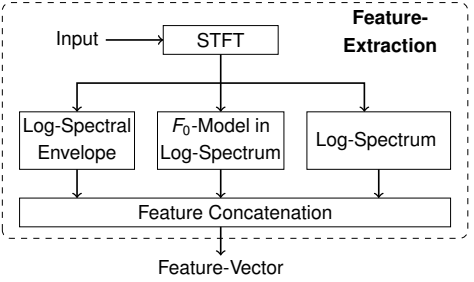
- Spectral envelope, F_0-model, Magnitude spectrum $\rightarrow$ Log-domain.
- F_0-model $\rightarrow$ largest cepstral coefficient i, and adjacent coefficients $i+1, i-1$

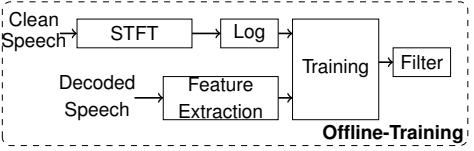
Offline-training

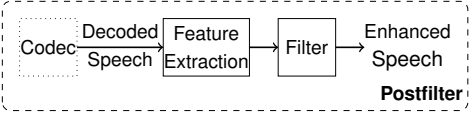
- Models in both frequency domain and perceptual domain investigated.

Postfilter

- Features computed from decoded signal.







Computational Complexity

Stage	Process	WMOPS
Features	Cepstrum (log-magnitude)	0.3099
	linear $\rightarrow$ log-magnitude	0.3099
Filtering	Matrix multiplication: **Ad**	17.163
Postprocessing	Magnitude$\rightarrow$Complex	0.0119
	log-magnitude$\rightarrow$linear	0.2980
Total		17.88

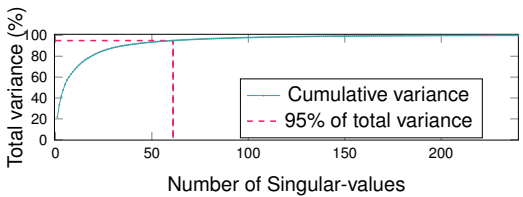


Figure: Plot of the cumulative variance over the singular values of filter **A**.

EVALUATION & RESULTS

Evaluation

- TIMIT dataset used for training and testing.
- Filter trained over 1000 speech samples $\rightarrow$ 340 female samples $\rightarrow$ randomly chosen from from training set.
- Gender specific filters in presented results.
- Codec $\rightarrow$ Similar to EVS in TCX mode used.
- Proposed postfilter applied on top of LTP-postfilter.

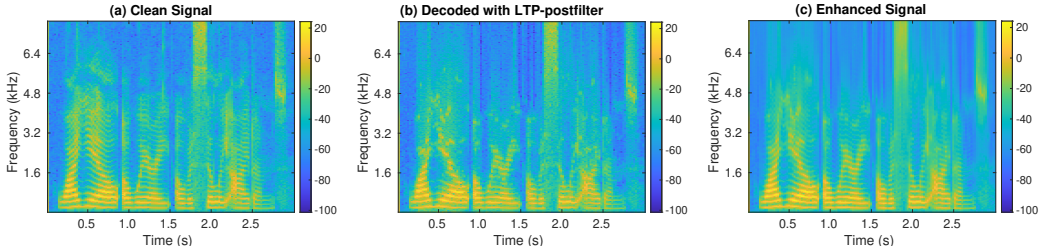


Figure: Spectrograms of a test sample: (a) Clean signal, (b) Decoded signal with LTP-postfilter, (c) Enhanced signal.

Objective Evaluation

Specifics

- 118 test samples (decoded speech) $\rightarrow$ 40 female, 78 male samples; randomly selected from TIMIT.
- Objective Measures:
  1. PSNR and ΔPSNR $\rightarrow$ signal-to-noise-ratio in the perceptual domain.
  2. PESQ and ΔPESQ
  3. POLQA and ΔPOLQA
- Evaluation results presented based on gender.

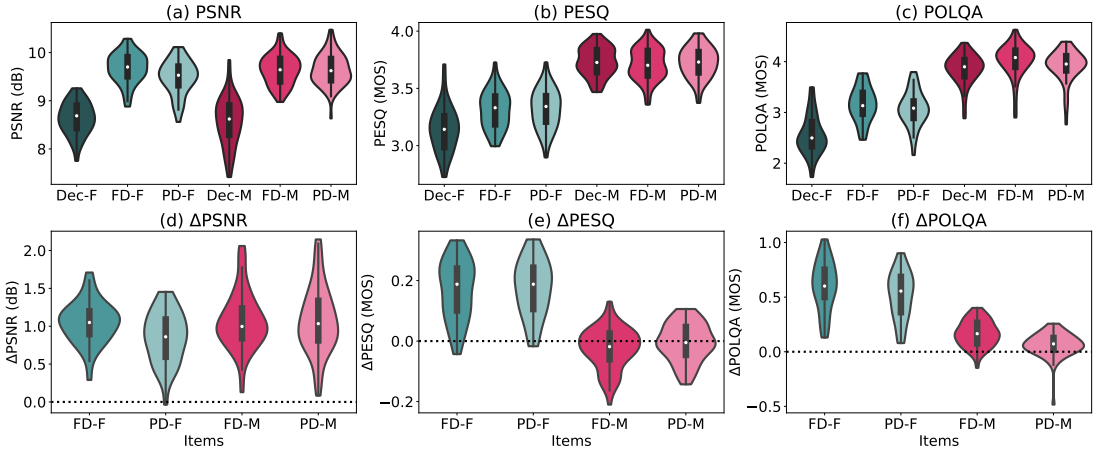
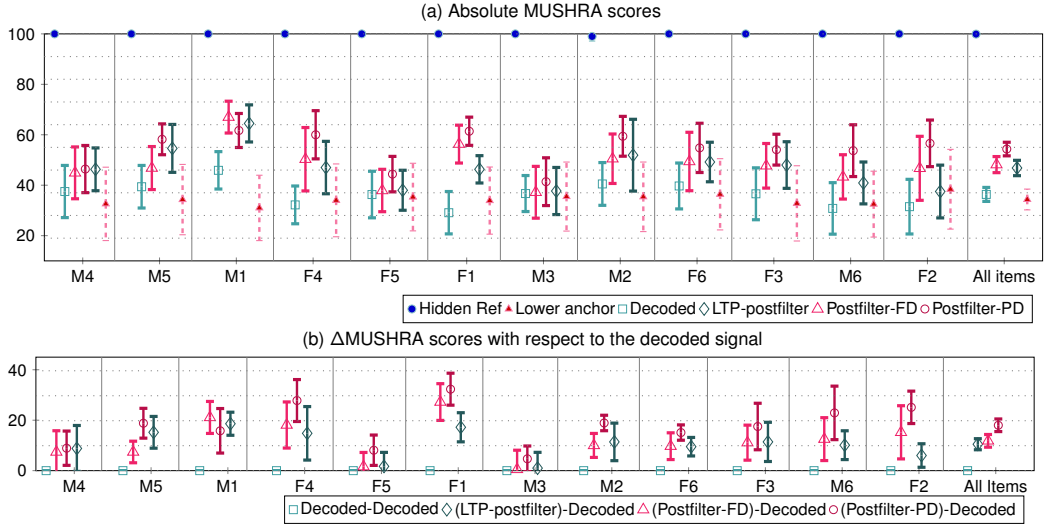


Figure: Distribution of the objective-measures via summary statistics and density trace of the absolute and Δ scores of the PSNR, PESQ, POLQA, for female (F) and male (M) samples in the Frequency domain (FD) and Perceptual domain (PD).

Subjective Evaluation: MUSHRA listening test

Specifics

- MUSHRA test with 10 listeners.
- 12 test items (6 female + 6 male) $\rightarrow$ 6 conditions $\rightarrow$ Lower-anchor, Hidden Reference, Decoded, LTP postfilter, Postfilter-FD, Postfilter-PD.
- 4 samples randomly chosen; 4 samples which showed the highest POLQA scores, 4 samples with least POLQA scores.


(a) Absolute MUSHRA scores
100
80
60
40
20
M4
M5
M1
F4
F5
F1
M3
M2
F6
F3
M6
F2
All items
Hidden Ref
Lower anchor
Decoded
LTP-postfilter
Postfilter-FD
Postfilter-PD
(b) ΔMUSHRA scores with respect to the decoded signal
40
20
0
M4
M5
M1
F4
F5
F1
M3
M2
F6
F3
M6
F2
All Items
Decoded-Decoded
(LTP-postfilter)-Decoded
(Postfilter-FD)-Decoded
(Postfilter-PD)-Decoded

Conclusion

- Speech coding $\rightarrow$ speech transmission.
- Postfilter for enhancement in speech and audio coding $\rightarrow$ incorporates information on harmonic structure.
- PSNR, PESQ, POLQA $\rightarrow$ positive improvement.
- Objective scores higher for females $\rightarrow$ accuracy of F_0-model.
- MUSHRA points higher for perceptual domain modelling.
- Postfilter successful in removing artefacts due to discontinuities.