ENHANCEMENT OF CODED SPEECH USING A MASK-BASED POST-FILTER

ICASSP 2020

Srikanth Korse, Kishan Gupta, Guillaume Fuchs

Fraunhofer

IIS

Overview

- Introduction
- Our Contribution
- Oracle Experiments
- Modified Signal Approximation
- Experimental Setup
- Results
- Conclusions

Introduction

- **CELP** based coding is integral part of state-of-the-art communication codecs such as **AMR-WB[1], 3GPP EVS[2]** etc.
- Quality of such codecs deteriorates at low bitrates due to high quantization noise.
- Usually, **post-filters** are employed to enhance the quality of coded speech at low bit-rates. These post-filters emphasize the pitch and formant structures of the coded speech using the **LPC and LTP** information.
- One such example of a post-filter is **G.718** [3].
- Recently a DNN-based post-filter in **cepstrum (Cepstrum-CNN) domain** was proposed in [4].

Our Contribution

- We propose a mask-based approach in spectral domain to enhance the quality of the coded speech.
- The proposed approach was implemented using:
  - Fully Connected Neural Network (**FCNN**)
  - Convolutional Encoder Decoder (**CED**)
  - Long Short Term Memory (**LSTM**)
- We compare our proposed system to heuristic post-filter adopted in the standard **G.718** and **Cepstrum-CNN**.
- The proposed model is trained on **single bitrate (6.65 kbps)** and tested on bitrates ranging from 6.65 kbps to 15.85 kbps.
- Robustness was validated by cross-database testing
- **POLQA**[5] and **MUSHRA**[6] was used for objective and subjective evaluation respectively.

Oracle Experiments (1/3)

- Spectral magnitude of enhanced speech ($|\hat{\boldsymbol{X}}(\boldsymbol{k},\boldsymbol{n})|$) is given by:

$$\left|\hat{X}(k,n)\right| = M(k,n) * |\tilde{X}(k,n)|$$

- If the mask is ideal, spectral magnitude of enhanced speech is same as spectral magnitude of clean speech.
- The ideal ratio mask (IRM) is given as: $\mathrm{IRM}(k,n) = \frac{|X(k,n)|}{|\tilde{X}(k,n)| + \epsilon}$

Mask Thresholds	6.65 kbps	8.85 kbps	12.65 kbps
[0,1]	38.94%	41.00%	44.09%
(1,2]	31.19%	33.44%	36.20%
(2,5]	21.40%	18.69%	14.66%
(5,∞]	8.46%	6.87%	5.05%

Table 1: Percentage of real-valued mask in different threshold regions measured at lowest three bitrates of AMR-WB.

Fraunhofer IIS

Oracle Experiments (2/3)

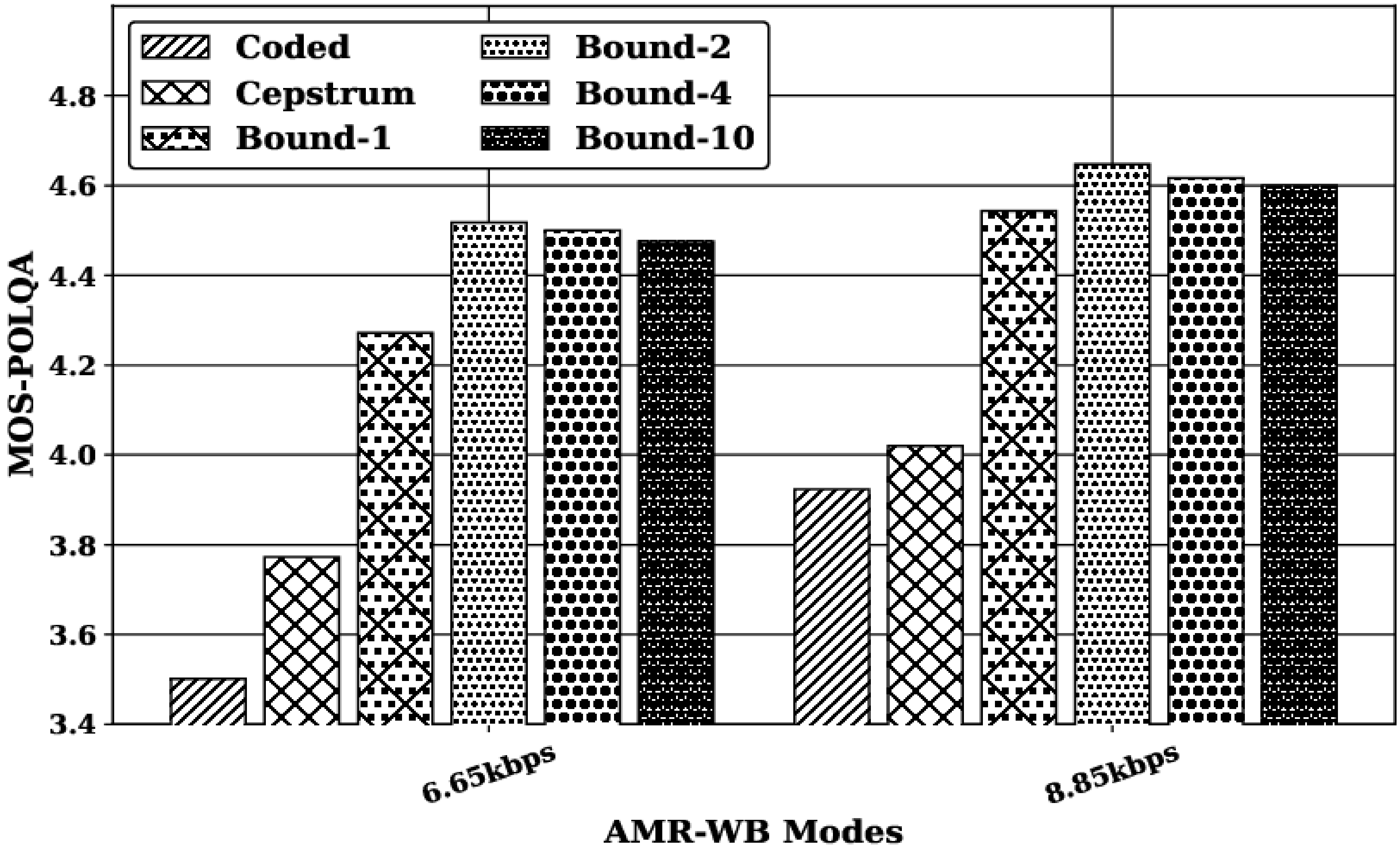
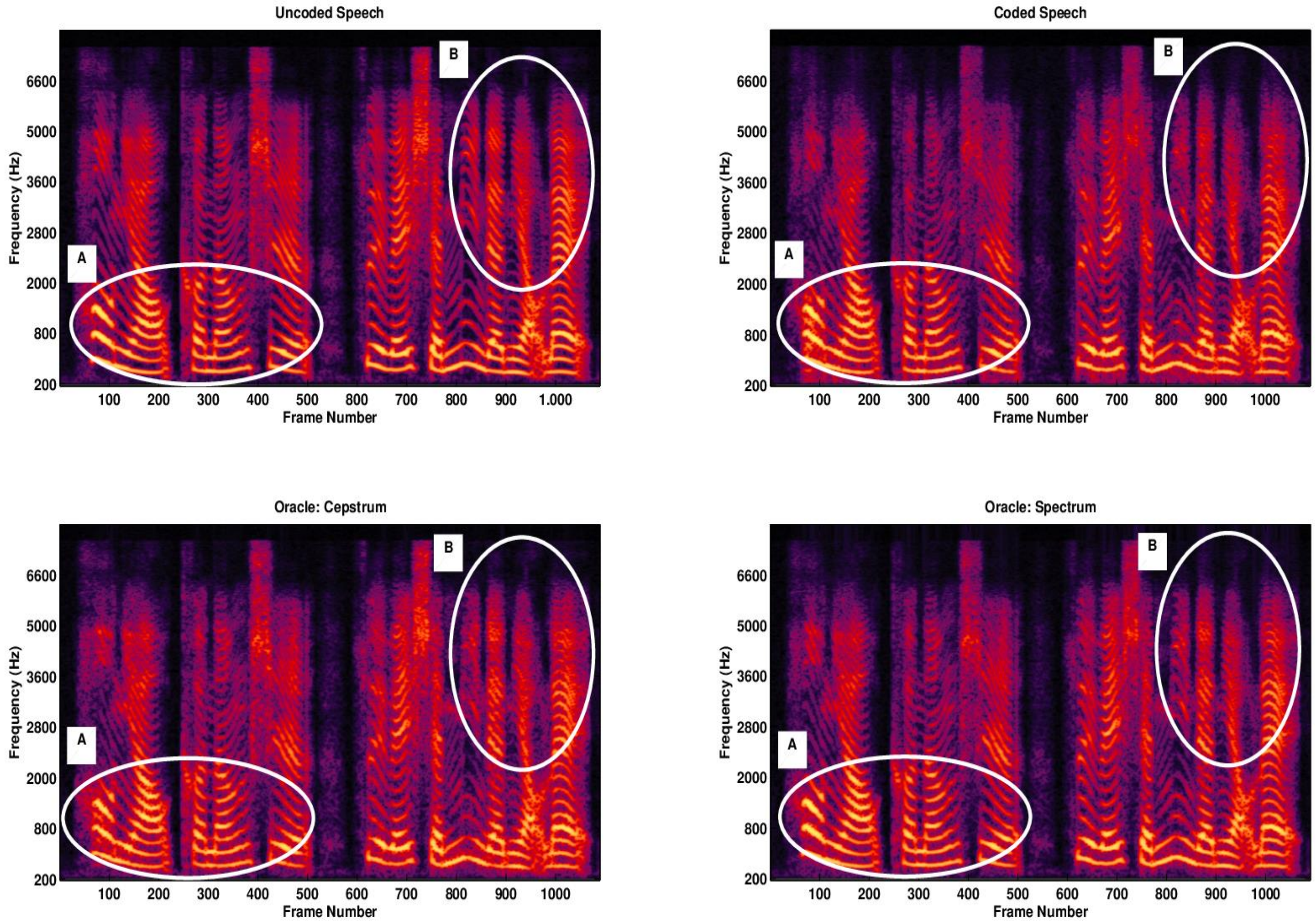


Fig 1: Average POLQA scores evaluating the oracle experiment at lowest 2 bitrates of AMR-WB (6.65kbps and 8.85kbps)

Oracle Experiments (3/3)



<u>Fig 2:</u> Spectogram comparison for Oracle case

Modified Signal Approximation

- The proposed model is trained using modified signal approximation (**mod-SA**)
- Main motivation behind mod-SA was to obtain generalized model.
- The main difference between our proposed mod-SA with the traditional signal approximation (**SA**) [7] are as follows:
  - The modified mask are computed as follows:

$$\widehat{M}(k,n) = \begin{Bmatrix} \mathrm{IRM}(k,n) & \text{if } \mathrm{IRM}(k,n) \leq \alpha \\ \rho & \text{if } \mathrm{IRM}(k,n) > \alpha \end{Bmatrix}$$

  - We set ρ as 1 and α as 2. This means, for bins, where IRM is greater than 2, the coded speech magnitude is kept unchanged.
  - The target is also modified as follows: $|\bar{X}(k,n)| = \widehat{M}(k,n) * |\tilde{X}(k,n)|$
  - The mean square error (mse) loss is computed between spectral magnitude of modified target and enhanced speech in log-magnitude domain.

Fraunhofer IIS

Experimental Setup (1/3)

Sampling Rate	16000
Transform	STFT
Analysis/Synthesis Window	Square root of Hann
Frame Size, Overlap	32ms, 50%
FFT Size	512
Processed Bandwidth	Upto 6.4kHz (205 bins)
DNN Input	Normalized Log Magnitude
Phase Processing	No

- The past frames were used as context frames
- Preprocessing for clean (target) speech: **P.341 filter[8]** (cutoff frequency 7kHz), active speech level was adjusted to -26 dBov [9].
- The coded speech was also post-processed with P.341 filter.

Experimental Setup (2/3)

- 3 models were tested:
  - **FCNN:**
    - Input Layer Size: 820 (3 past frames and current frame).
    - 2 hidden layers with 1024 units and ReLU activations.
    - Batch normalization, dropout of 0.1.
  - **LSTM:**
    - 2 LSTM layers with 400 and 205 units respectively
    - Input: 10 time steps (9 past frames and current frame).
    - A dropout of 0.1 and recurrent dropout of 0.2 is used.
  - **CED:**
    - Input: 6 time steps (5 past frames plus current one)
    - ELU activation, batch normalization, skip connections.
- Output: 205 units of sigmoid activations with scaling factor of 2.

Experimental Setup (3/3)

Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Convergence/Epochs	Early Stopping
Training/ Validation	NTT-AT [10] *
Testing	NTT-AT*/ TIMIT[11]

*All files were downsampled to 16kHz and a passive mono downmix was obtained.

Fraunhofer IIS

Results (1/6)

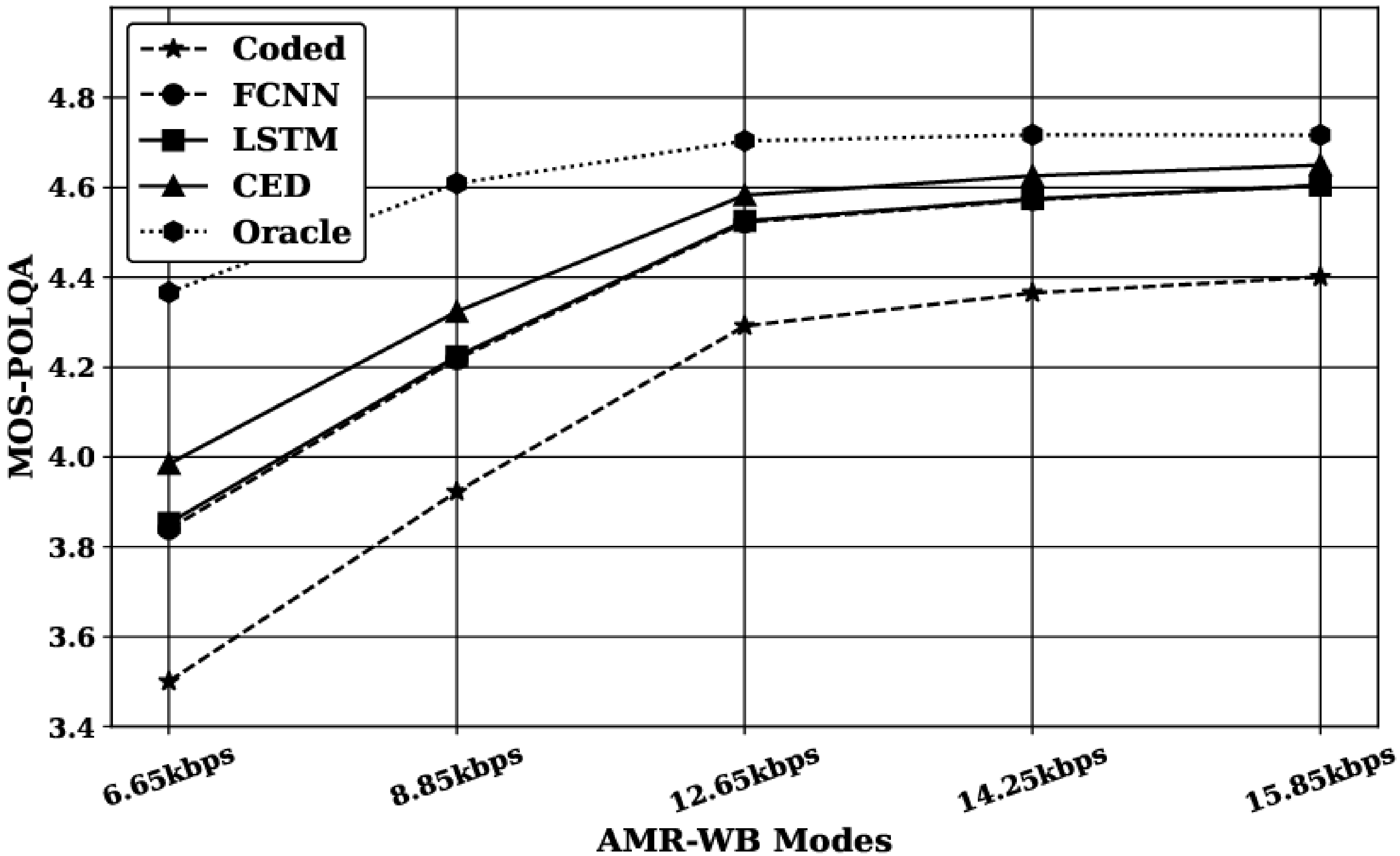


Fig 3: POLQA scores evaluating the performance of the FCNN, LSTM and CED architectures using the NTT test set (lowest 5 modes of AMR-WB).

Fraunhofer IIS

Results (2/6)

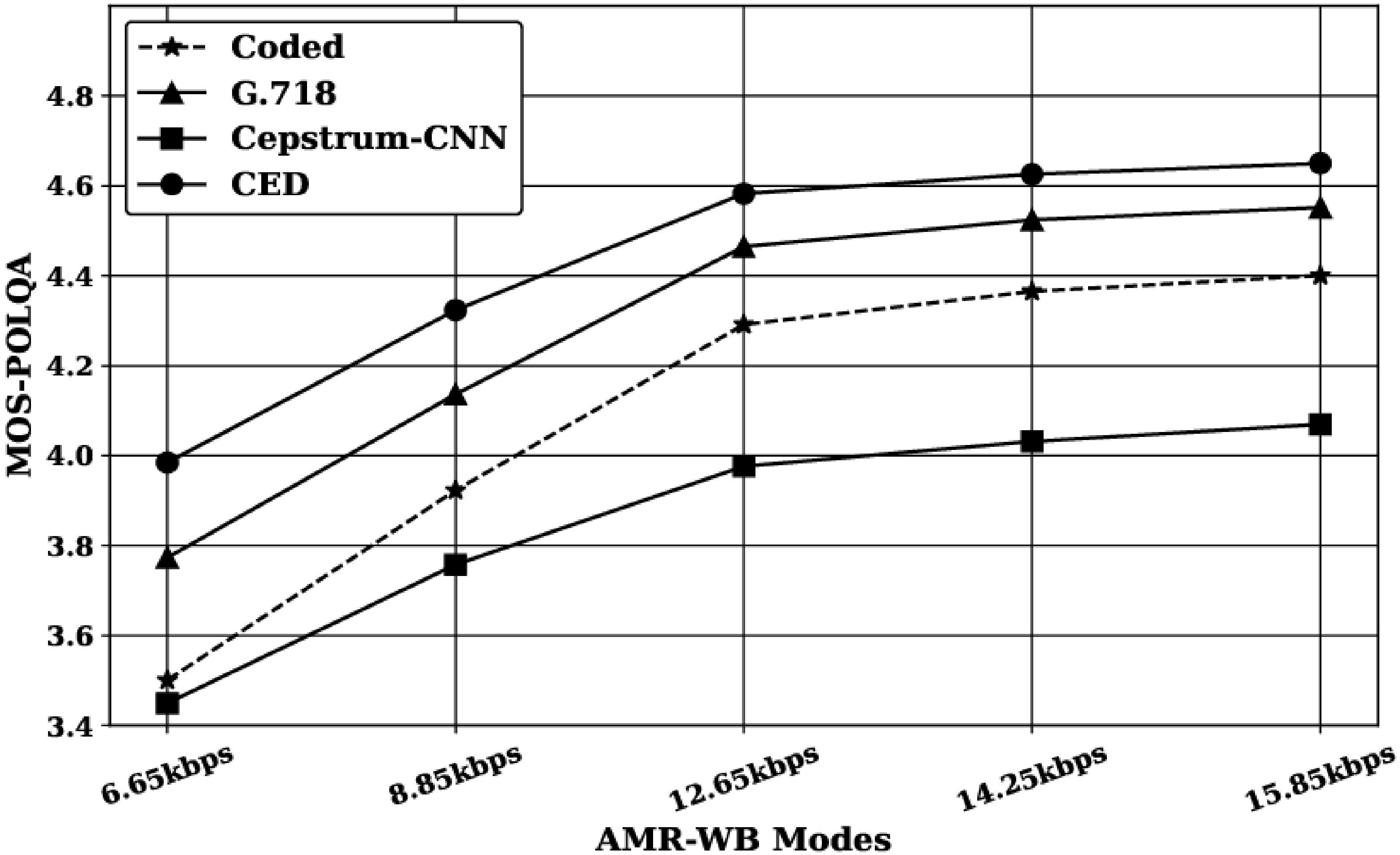


Fig 4: POLQA scores evaluating the performance of the Cepstrum-CNN, CED and G.718 using the NTT test set (lowest 5 modes of AMR-WB).

Fraunhofer IIS

Results (3/6)

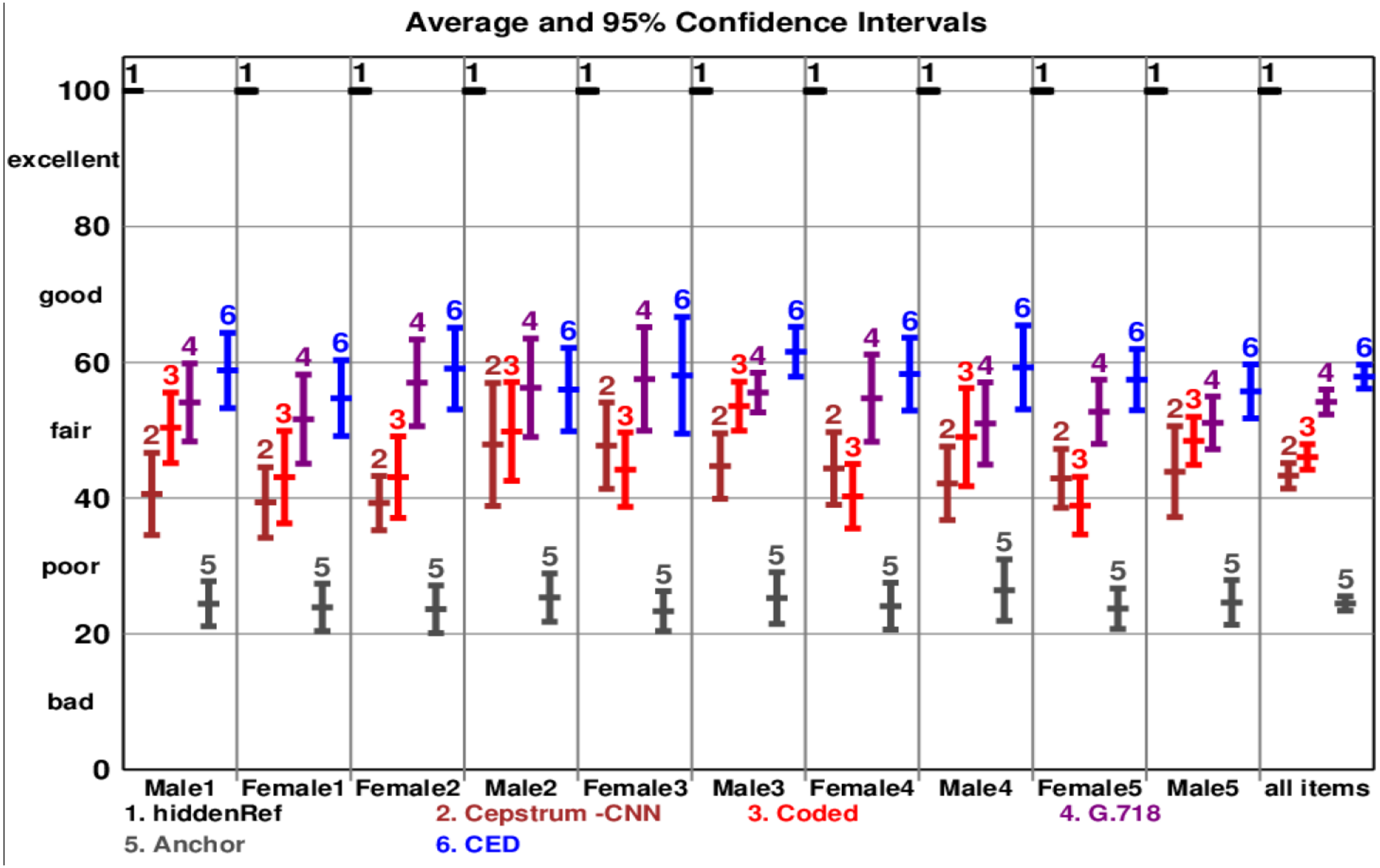


Fig 5: Average MUSHRA scores of 11 listeners at 6.65 kbps.

Fraunhofer IIS

Results (4/6)

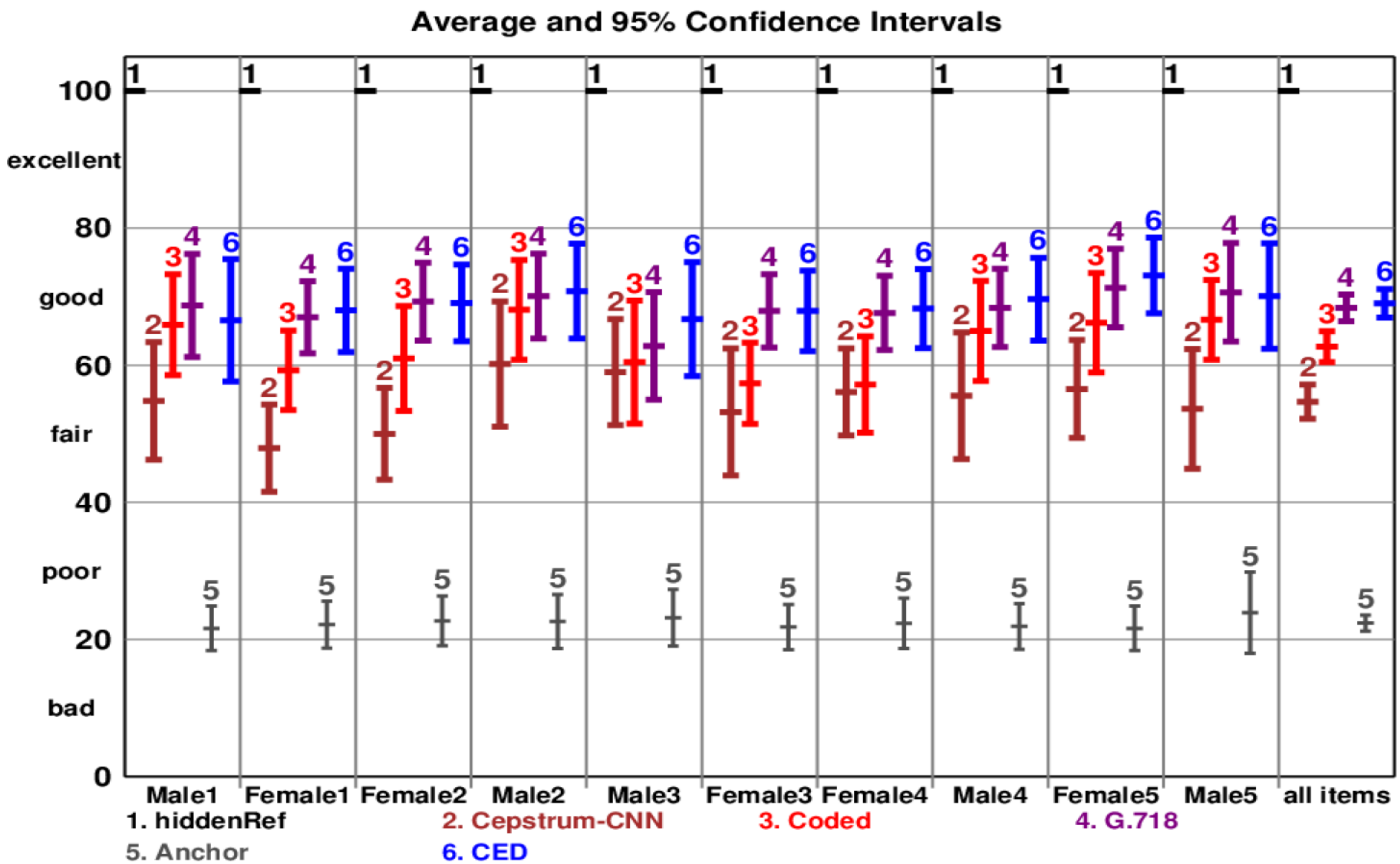
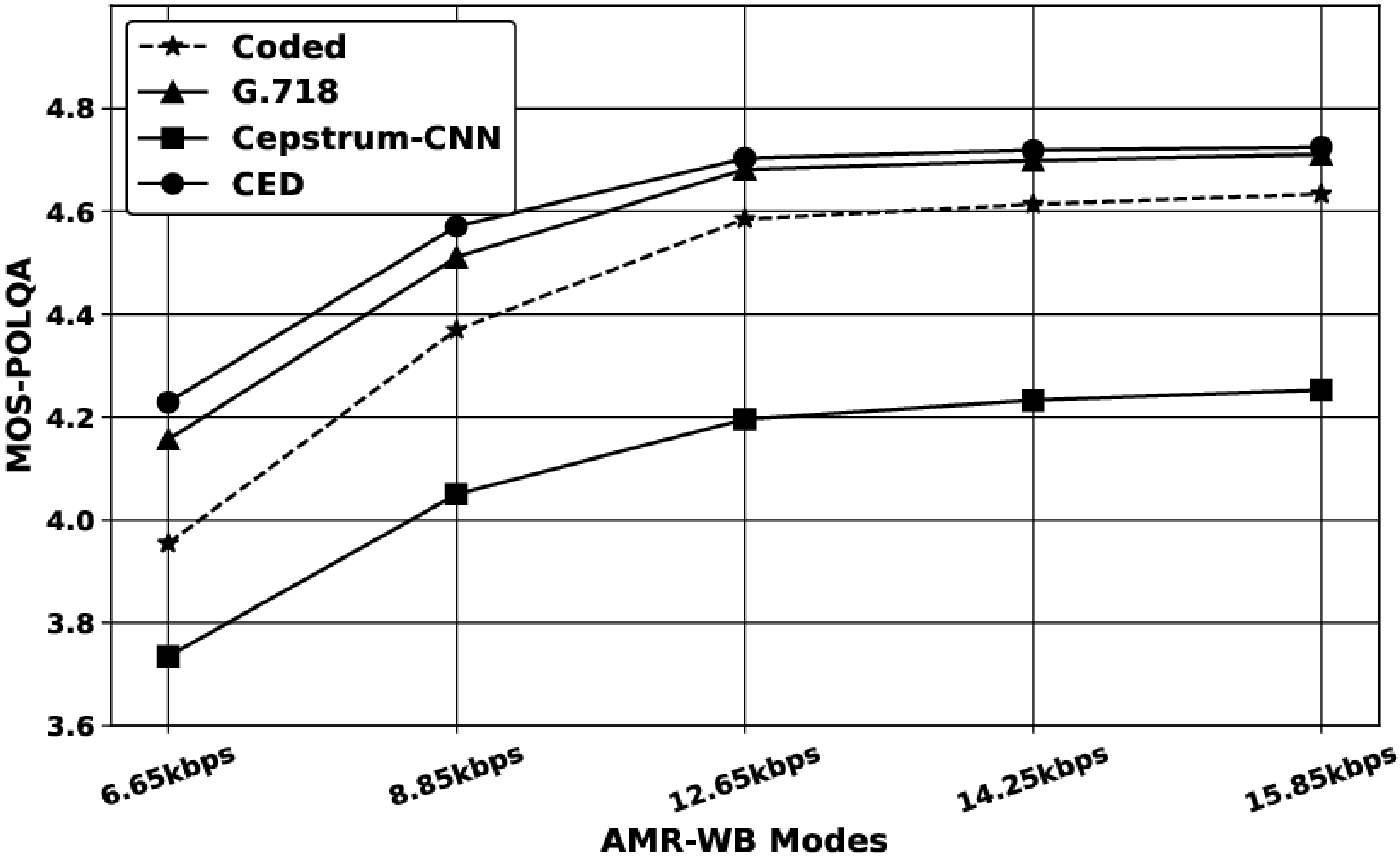


Fig 6: Average MUSHRA scores of 11 listeners at 12.65 kbps.

Results (5/6)



<u>Fig 7:</u> POLQA scores evaluating the performance of Cepstrum-CNN, CED and G.718 using the TIMIT test set (lowest 5 modes of AMR-WB).

Results(6/6)

Network Architecture	Number of Parmaters	Frame Size
FCNN	2,108,621	32 ms
LSTM	1,468,120	32 ms
CED	**147,292**	32 ms
Cepstrum-CNN	419,805	**20 ms**

Table 2: Comparison of the number of parameters in different network architectures.

Conclusion

- We have proposed convolutional encoder-decoder (CED) based post-filter for enhancing the perceptual quality of coded speech.
- Our proposed post-filter makes no assumption of signal or noise characteristics.
- The proposed post-filter estimates a real valued mask per time-frequency bin.
- The post-filter is trained using modified signal approximation (mod-SA) in order to obtain generalized model.
- The generalized model works well at even higher bitrates inspite of being trained on lowest bitrate.
- Robustness of our proposed model is proved by cross-database testing.

Fraunhofer
IIS

References(1/2)

- [1] 3GPP, “Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Transcoding functions,” 3rd Generation Partnership Project (3GPP), TS 26.190, 12 2009. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/26190.htm
- [2] ——, “TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12),” 3rd Generation Partnership Project (3GPP), TS 26.445, 12 2014. [Online]. Available: http://www.3gpp.org/ftp/Specs/htmlinfo/26445.htm
- [3] ITU-T Recommendation G.718, “Frame error robust narrowband and wideband embedded variable bit-rate coding of speech and audio from 8–32 kbit/s,” 2008.
- [4] Z. Zhao, H. Liu, and T. Fingscheidt, “Convolutional neural networks to enhance coded speech,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 4, pp.663–678, April 2019.
- [5] Perceptual objective listening quality assessment (POLQA), ITU-T Recommendation P.863, 2011. [Online]. Available: http://www.itu.int/rec/T-REC-P.863/en

References(2/2)

- [6] Recommendation BS.1534, Method for the subjective assessment of intermediate quality levels of coding systems, ITU-R, 2003.
- [7] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Dec 2014, pp. 577–581.
- [8] ITU-T G.191, "Software tools for speech and audio coding standardization," 2005.
- [9] ITU-T P.56, "Objective measurement of active speech level," 2011.

Fraunhofer IIS