NON-INTRUSIVE SPEECH QUALITY ASSESSMENT USING NEURAL NETWORK

Anderson R. Avila[1], Hannes Gamper[2], Chandan Reddy[3], Ross Cutler[3], Ivan Tashev[2], Johannes Gehrke[3]

[1]Institut National de la Recherche Scientifique, Montreal, QC, Canada
[2]Microsoft Research Labs, Redmond, WA, USA
[3]Microsoft Corporation, Redmond, WA, USA

Introduction

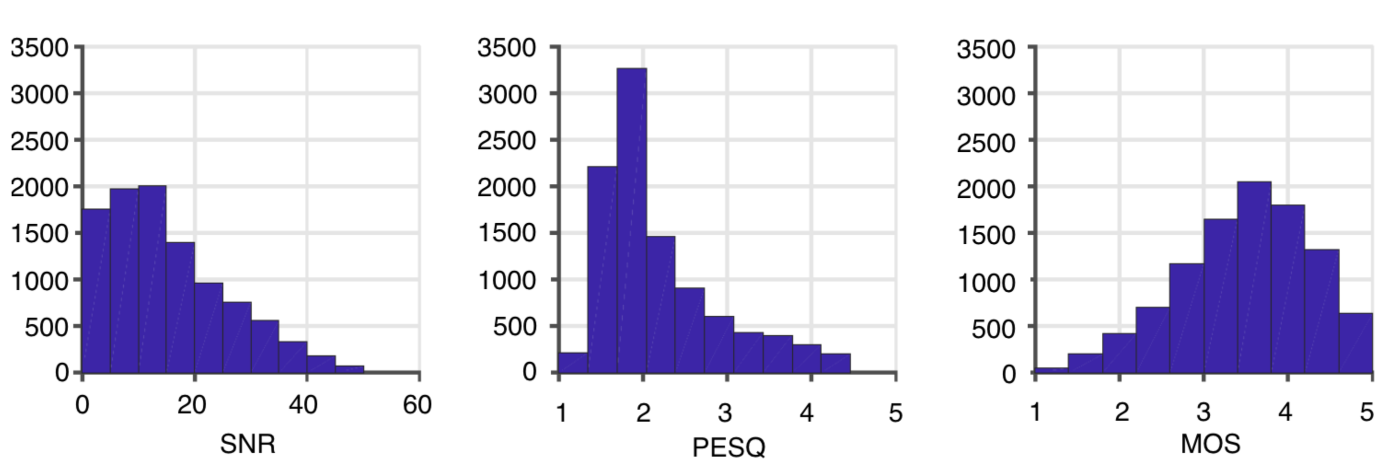
Estimating the speech quality as perceived by humans is a challenging but important task for many multimedia applications. In this work, three neural network-based approaches are proposed to estimate the mean opinion score (MOS) of reverberant speech in background noise.

Contribution

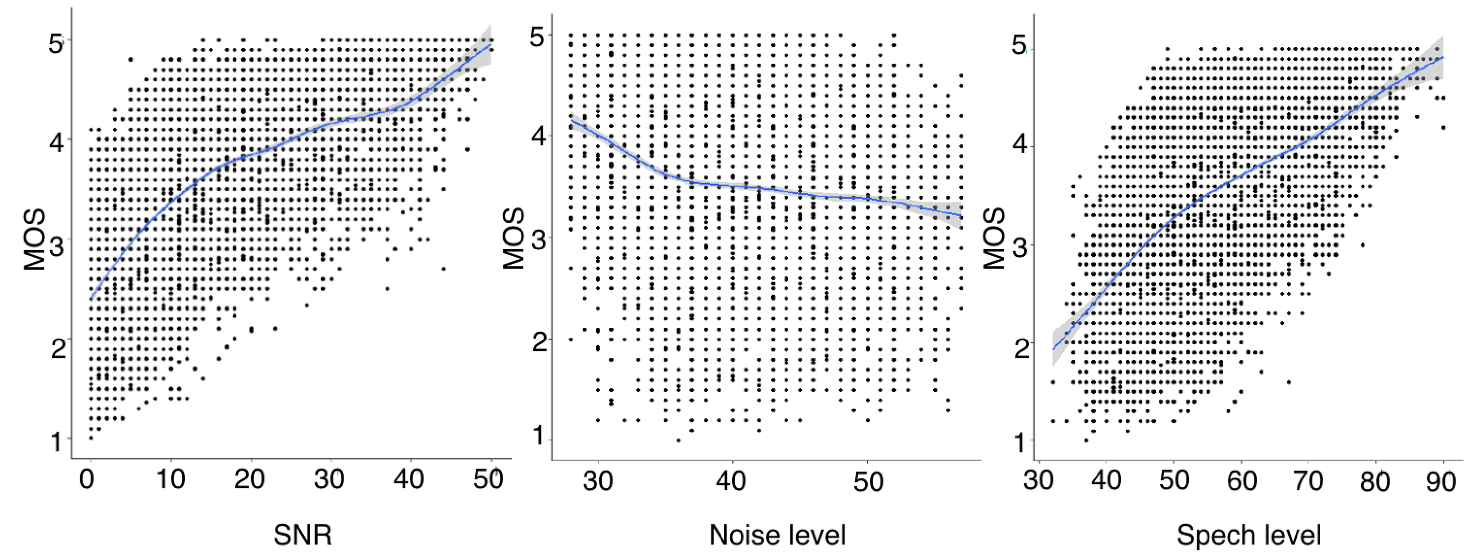
- Challenging speech scenarios with various distortions (reverberation, ambient noise, speech enhancement)
- Three neural network-based approaches for estimating MOS
- Proposed approaches outperformed three benchmarks (PESQ, SRMR and P.563)

Data Generation

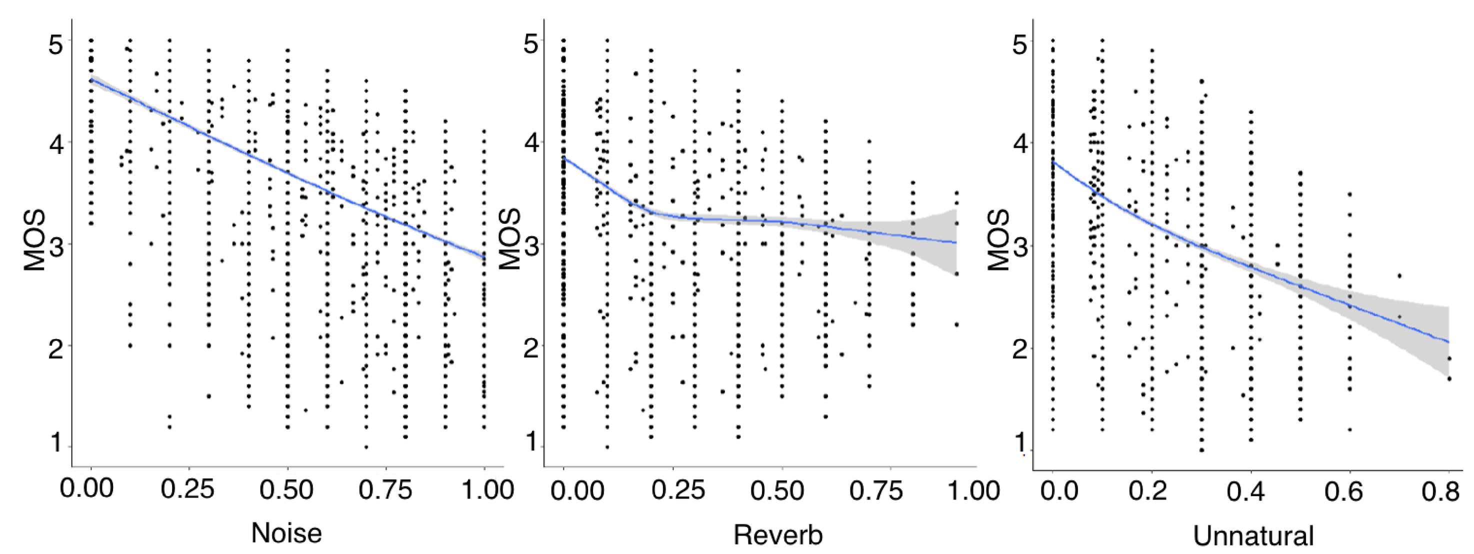
- 10.000 speech samples (male, female and children)
- 90 dB FS (Clipping)
- Speech level with mean of 65 dB and deviation of 8 dB
- Noise level with mean of 45 dB SPL and deviation of 15 dB
- 120 Room impulse response (RIR) randomly selected
- RT_{60} ranging from 300 to 500 ms
- Speakers and microphones distance varying between 0.5 and 3 meters
- Anechoic and close-talk microphone also included
- Offices (80 %), homes (10 %) and others (10 %)
- Labels are attained using crowd-sourcing



Exploratory Data Analysis



MOS versus actual SNR, noise and speech levels

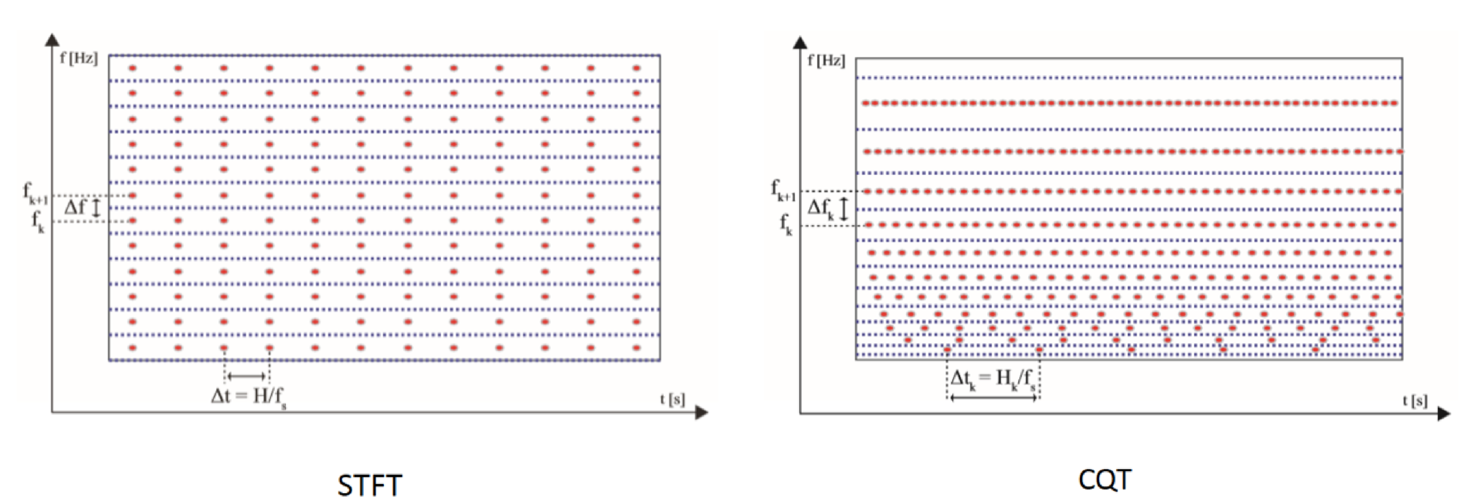


MOS vs noise, reverberation and unnaturalness as perceived by raters

Proposed Approaches

1) Constant Q Spectral + CNN

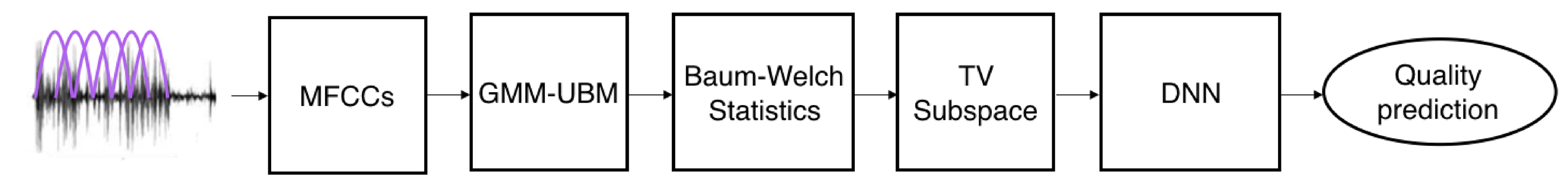
$$Q_c = \frac{f_c}{\delta_f}$$



Source: Todisco, Massimiliano, Héctor Delgado, and Nicholas Evans. "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients." *Speaker Odyssey Workshop, Bilbao, Spain.* Vol. 25. 2016.

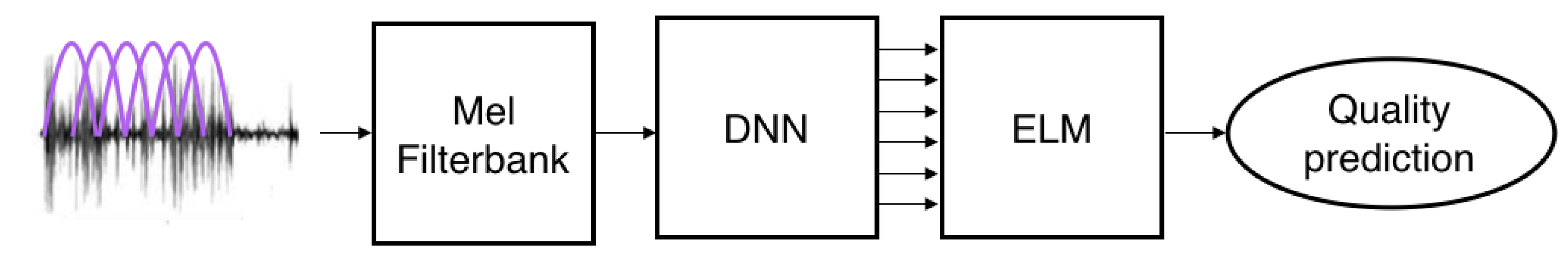
Perceptually motivated, the constant Q transform (CQT) allows better time-frequency resolution by applying a quality factor across different frequencies.

2) I-vector + DNN



We investigate the use of i-vectors as input to a DNN model. Known for capturing both speaker and channel variability, the framework maps into the total variability space (TV) a list of feature vectors, $O = \{o_t\}_{t=1}^{N}$, where $o_t \in \mathbb{R}^F$, and N is the frame index, into a fixed-length vector, $n \in \mathbb{R}^D$.

3) Mel-Frequency + DNN + Extreme Learning Machine



We explore utterance-level features obtained as result of statistics extracted from segment-level representation provided by a DNN model. These utterance-level features are used as input to an efficient single-hidden-layer neural network, known as extremely learning machine (ELM), to predict speech quality.

Results

Model	ρ	MSE
PESQ	0.70	0.25
SRMR	0.60	0.31
P.563	0.55	0.36
Constant Q (Spectrum) + CNN	0.72	0.30
i-vector + DNN	0.78	0.22
Mel-Frequency + DNN	0.86	0.18
Mel-Frequency + DNN + ELM	**0.87**	**0.15**

Future work

As future work, we will evaluate the proposed methods on an extended dataset with network impairments. We will also consider training a DNN model using the raw signal.