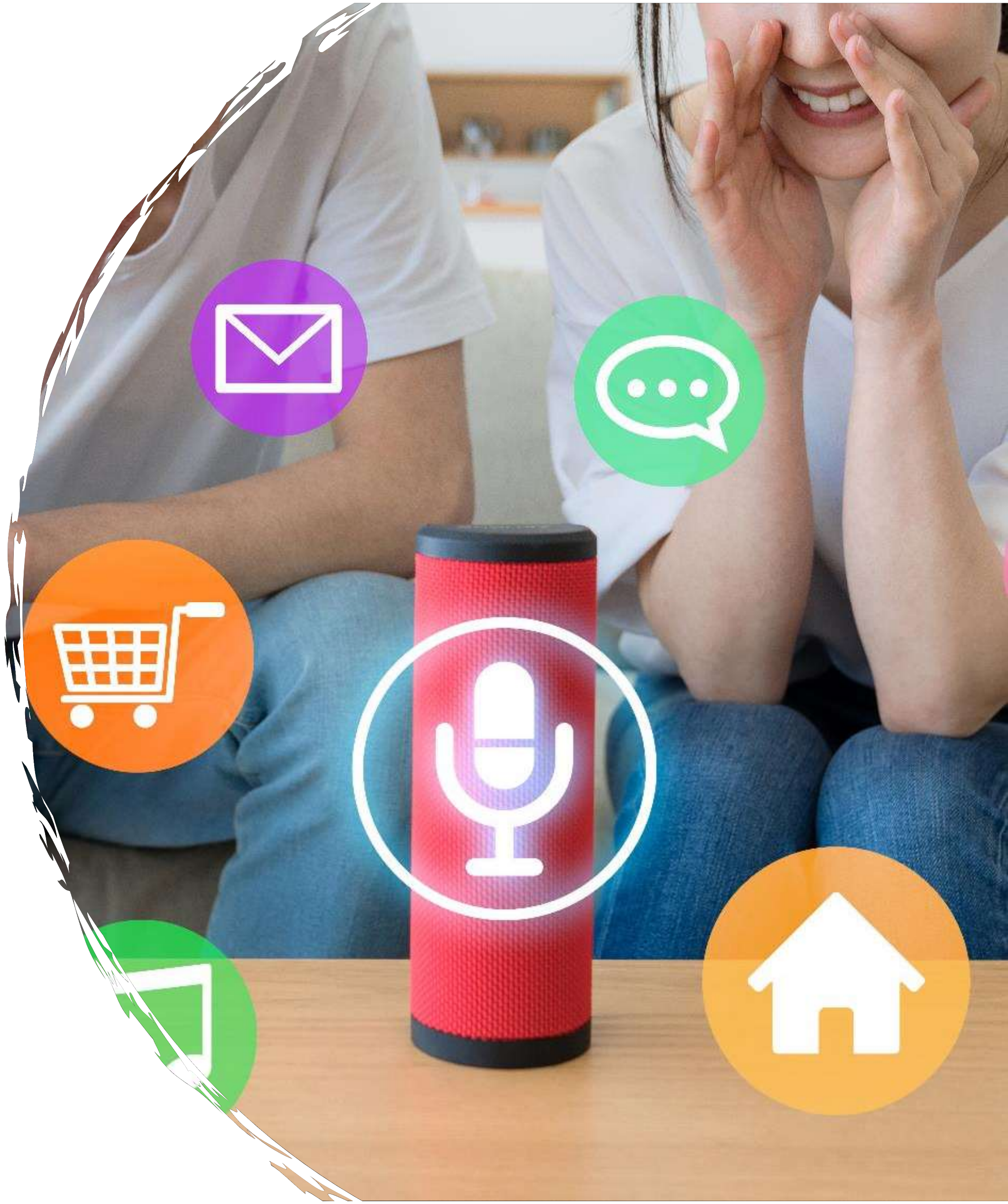
STQ Workshop

Recent Advances in Non-Intrusive Speech Quality Prediction

Wafaa Wardah

21/11/2022



Speech Quality

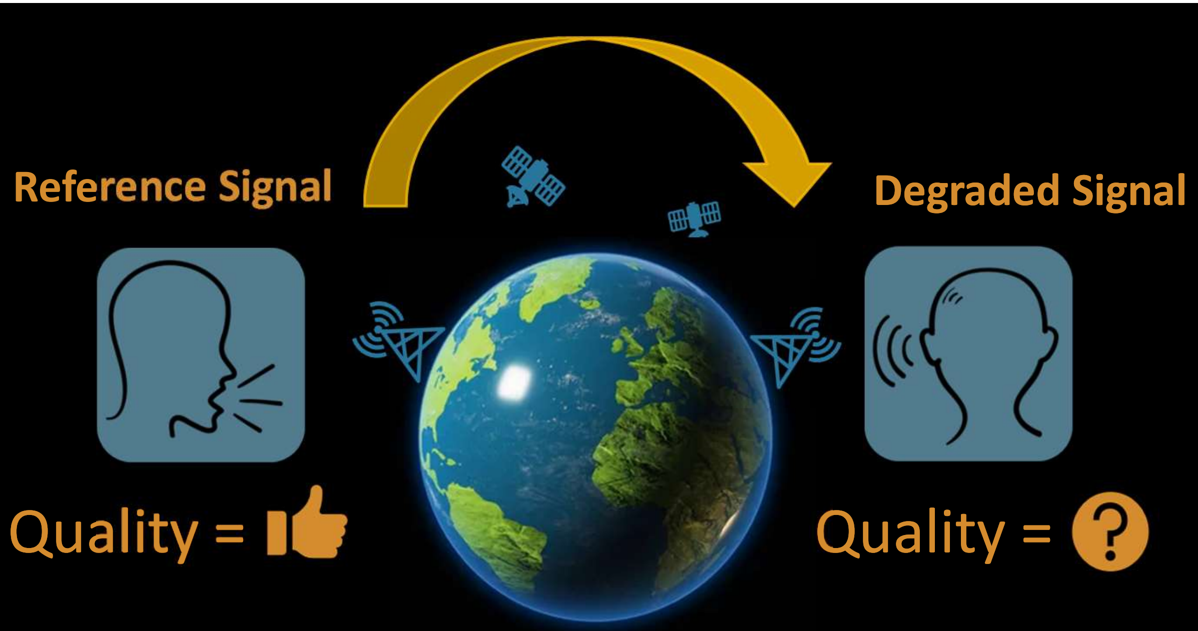
Quality Prediction

ML Model NISQA

Standardization

Ongoing Work

Highlights



Reference Signal

Degraded Signal

Quality = 👍

Quality = ?

- Clean reference signal is sent over networks via equipment that inevitably leads to degradation.
- A good way to measure the quality of the transmitted speech is in terms of how humans on the receiving perceive it.
- Non-intrusive (single-ended) quality assessment means using only the degraded signal for monitoring the quality, whereas an intrusive (double-ended) approach is having access to the clean reference signal as well.

Speech Quality

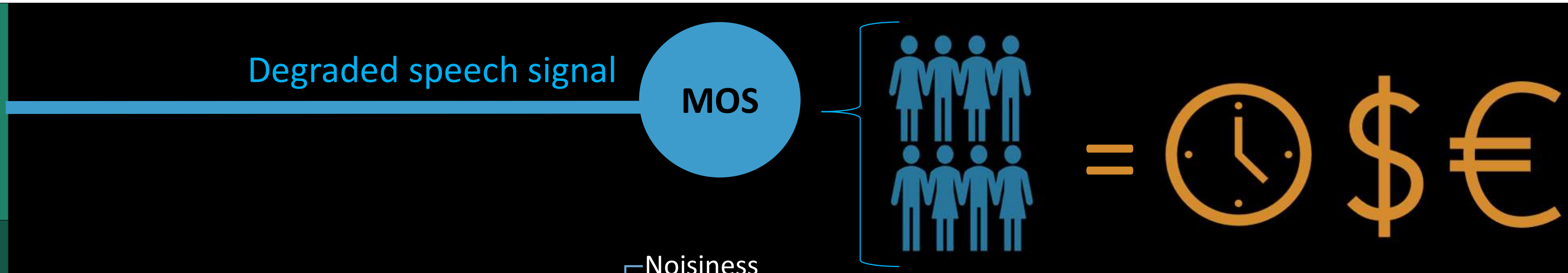
Quality Prediction

ML Model NISQA

Standard-ization

Ongoing Work

Highlights



Dimensions
- Noisiness
- Coloration
- Discontinuity
- Loudness

Absolute Category Rating (ACR) scale
1 – Bad
2 – Poor
3 – Fair
4 – Good
5 – Excellent

- Traditionally, we derive this perceived quality from subjective listening tests.
  - ITU-T Recommendation P.800 Methods for subjective determination of transmission quality carried out in a controlled lab environment.
  - ITU-T Recommendation P.808 Subjective evaluation of speech quality with a crowdsourcing approach (Toolkit)
- Unfortunately, listening tests are costly, time-consuming, inconvenient, and not portable, so instrumental models that can automatically predict speech quality have been developed.

Speech Quality

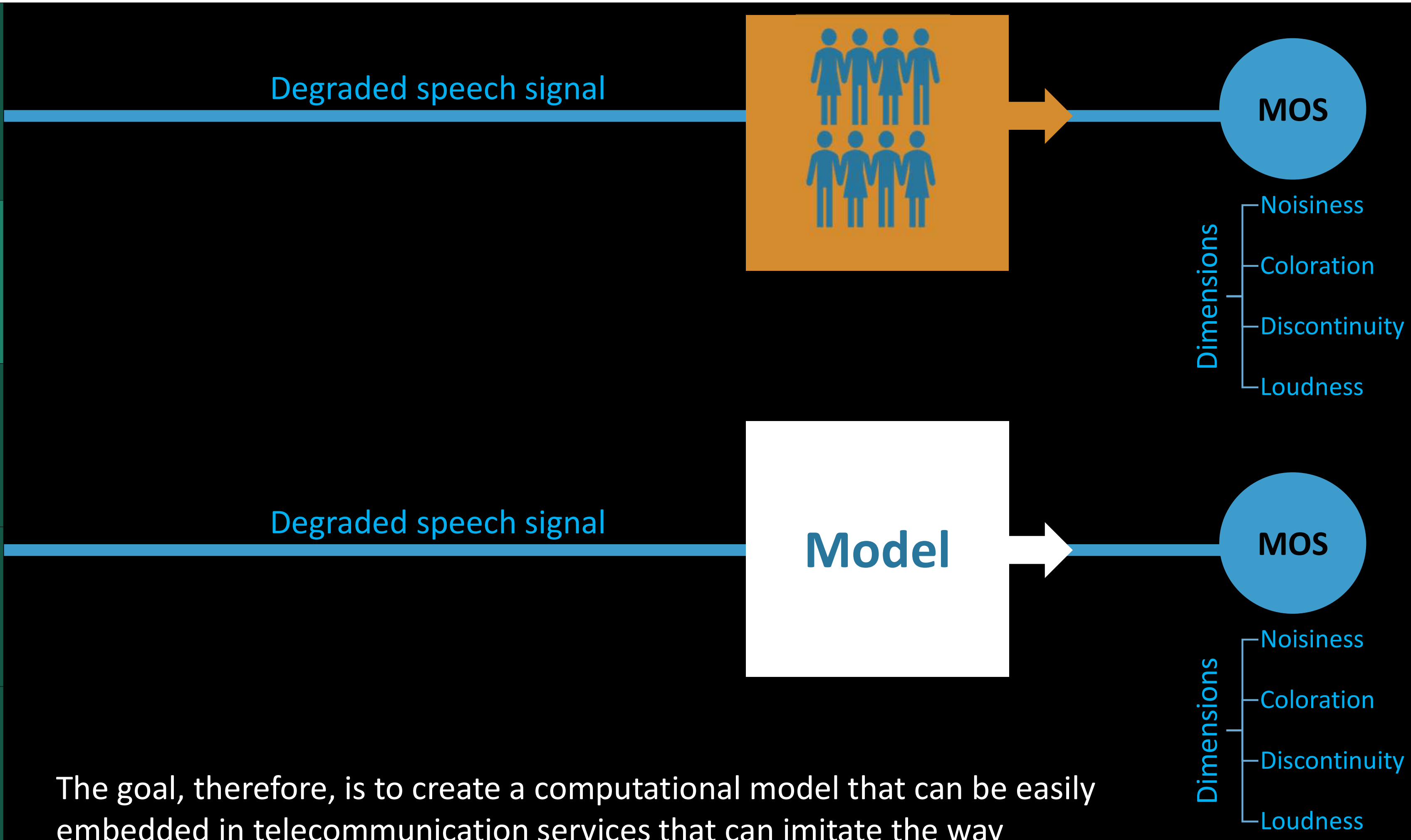
Quality Prediction

ML Model NISQA

Standard-ization

Ongoing Work

Highlights



The goal, therefore, is to create a computational model that can be easily embedded in telecommunication services that can imitate the way human participants perceive and therefore rate the overall quality as well as the quality dimensions.

Speech Quality

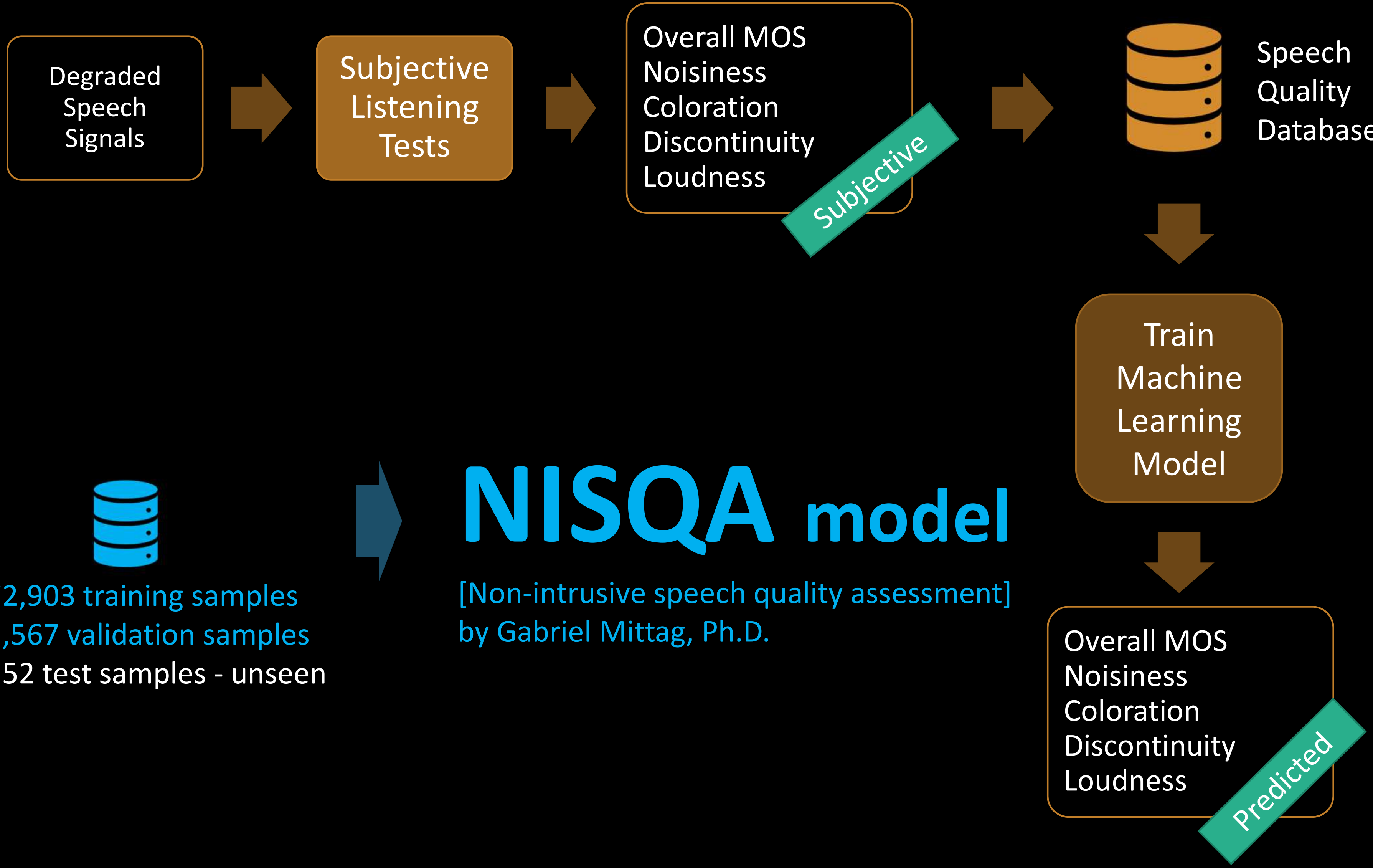
Quality Prediction

ML Model NISQA

Standard-ization

Ongoing Work

Highlights



Degraded Speech Signals

Subjective Listening Tests

Overall MOS
Noisiness
Coloration
Discontinuity
Loudness

Subjective

Speech Quality Database

Train Machine Learning Model

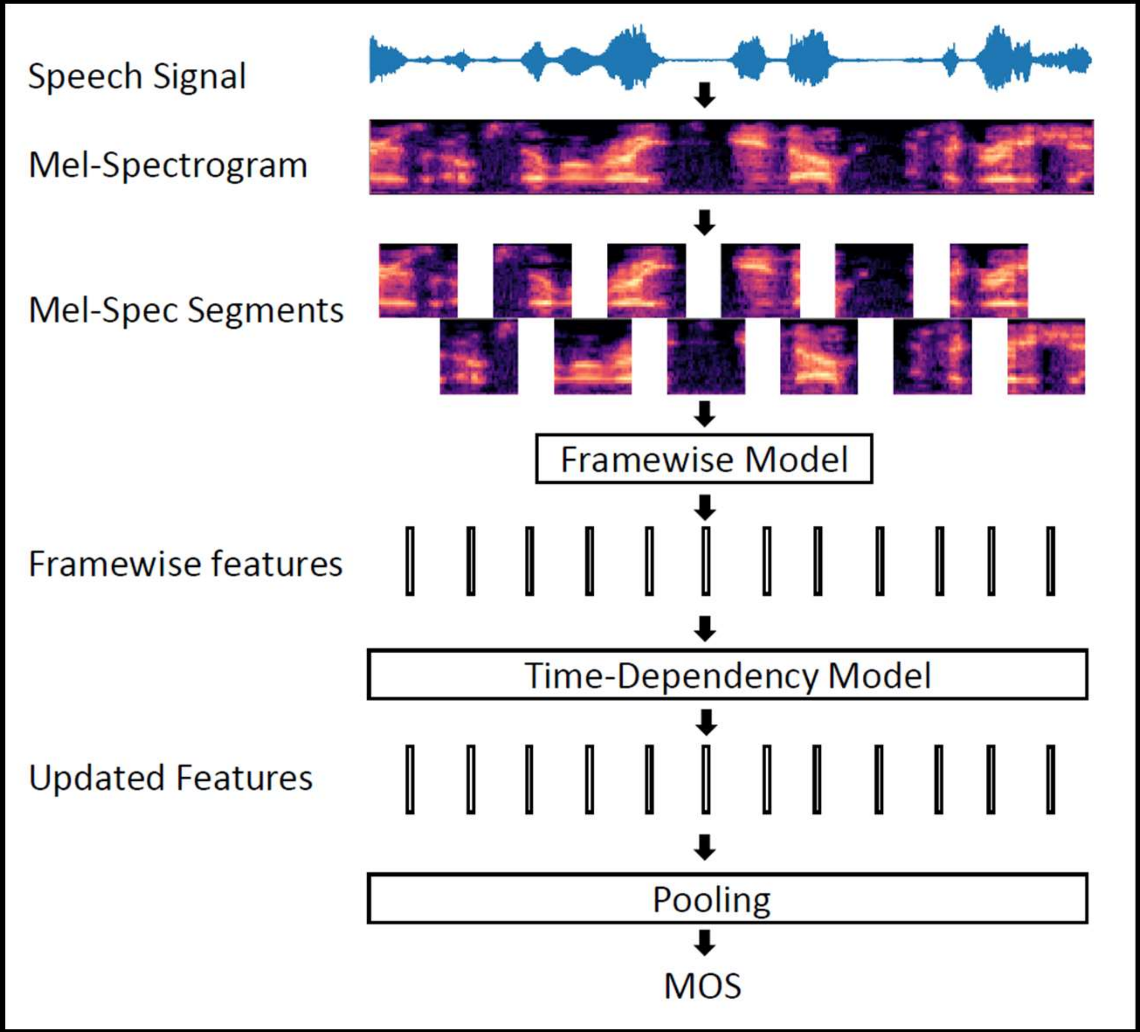
Overall MOS
Noisiness
Coloration
Discontinuity
Loudness

Predicted

72,903 training samples
9,567 validation samples
952 test samples - unseen

NISQA model

[Non-intrusive speech quality assessment]
by Gabriel Mittag, Ph.D.



10-second speech signal
=
250 segments
with 73% overlap

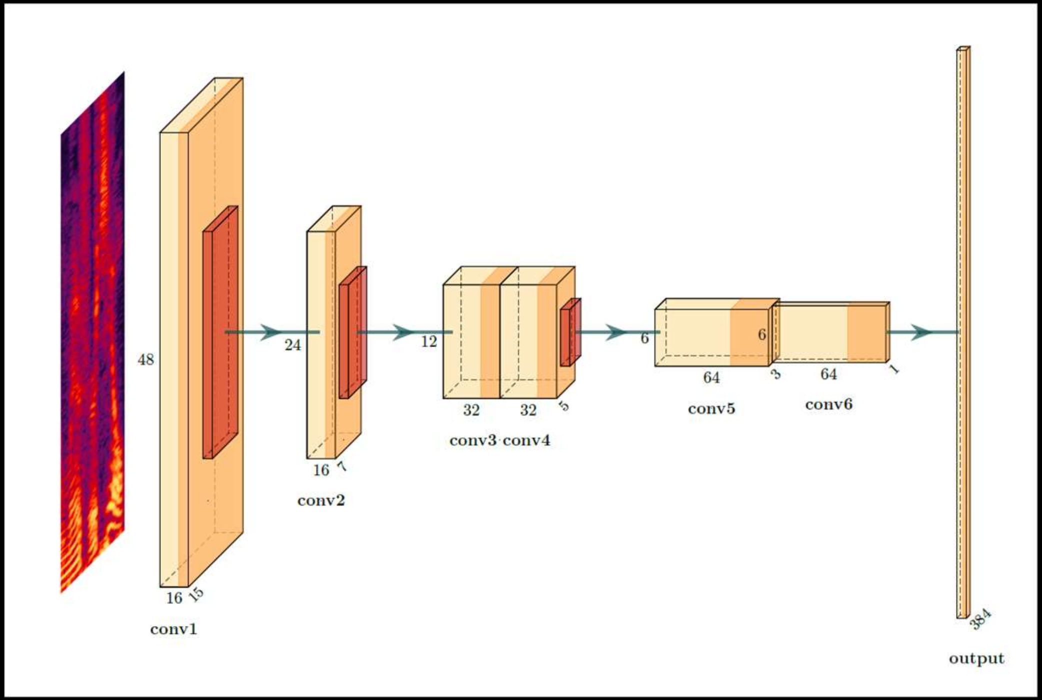
- The input to the model is a speech signal.
- The output is the predicted ratings for:
  - overall quality MOS
  - noisiness
  - loudness
  - coloration
  - discontinuity

Framewise model

- CNNs are most commonly used in the field of image classification and have the ability to learn a suitable set of features

- While an RGB image has three channels – one for each color – the Mel-spec input has only one channel, representing the spectrogram's amplitude.

- convolutional neural network
- 6 convolutional layers
- 3 max-pooling layers
- Flattened output of length 384

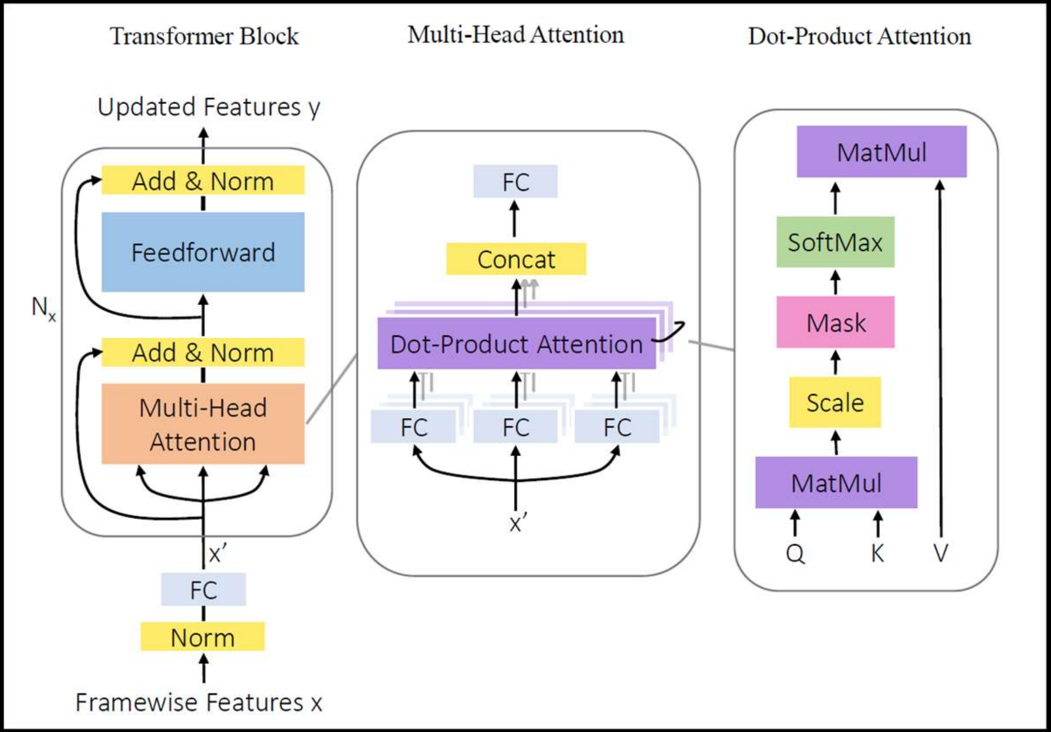


Time-dependency model

- Self-Attention network
- Based on the transformer encoder
- Single head, depth of 2 blocks

- The input to the Transformer block is the sequence of framewise features produced by the framewise model.

- It learns the temporal dependencies in the framewise features.

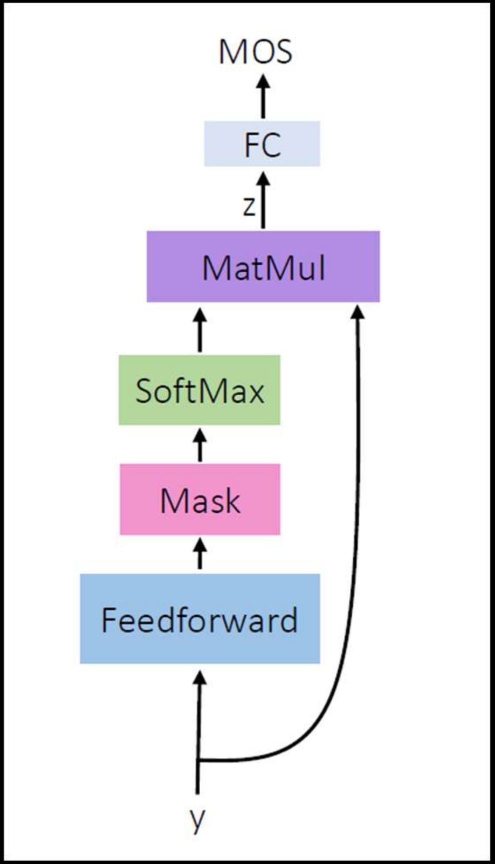


Pooling model

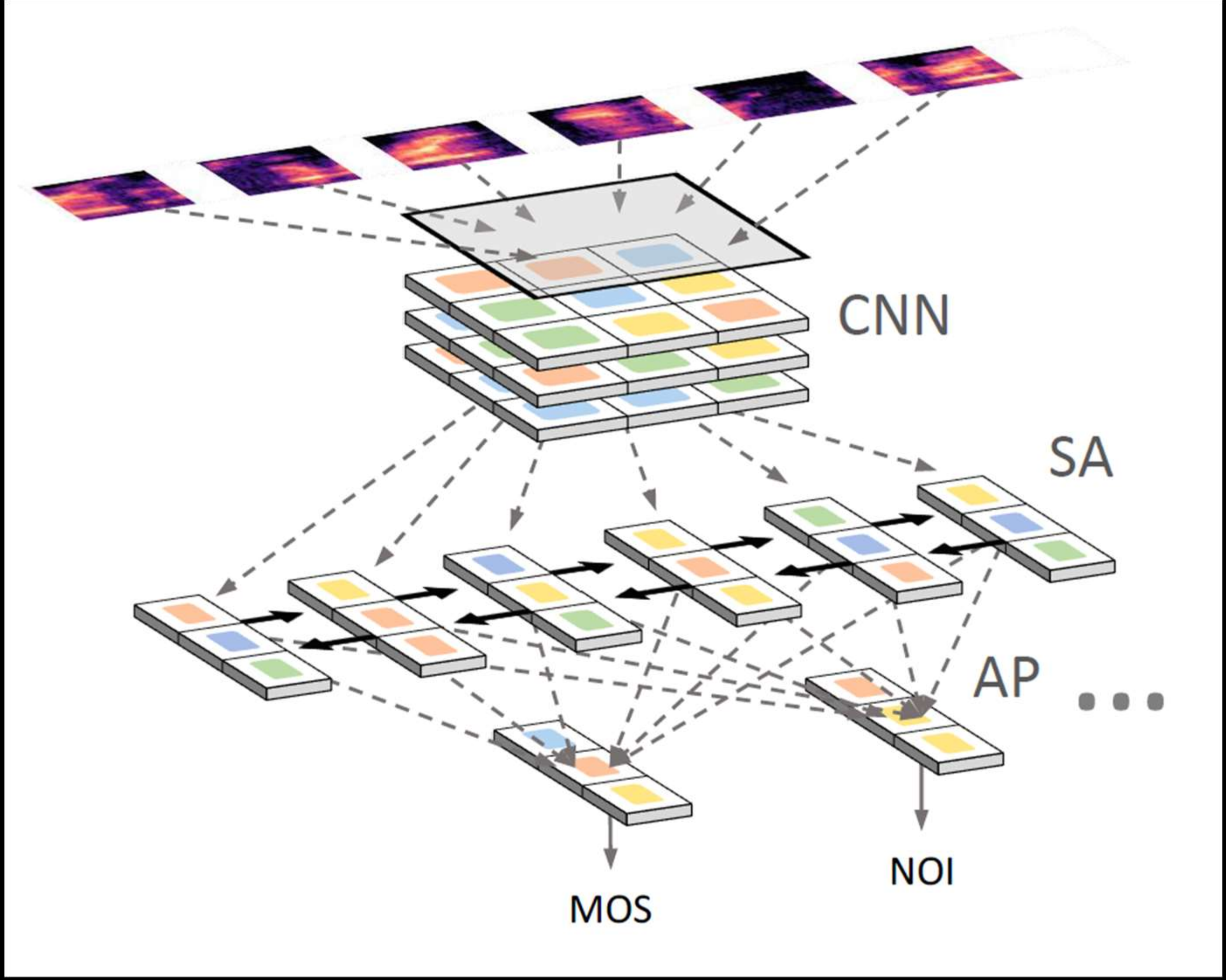
- Attention-pooling

- The input to the Attention-pooling block is the output matrix containing time domain information produced by the Self-Attention model.

- The final output produced by this Attention-pooling model is the predicted overall quality MOS, noisiness, loudness, coloration, and discontinuity scores.



- Multitask problem – five scores to predict
- Mel-spec features are calculated by the same CNN and Self-Attention network for each dimension
- CNN and Self-Attention network is shared across all tasks
- Outputs of each Self-Attention time-step are then the input for five individual pooling blocks that predict the overall MOS and the dimension scores.



Speech Quality

Quality Prediction

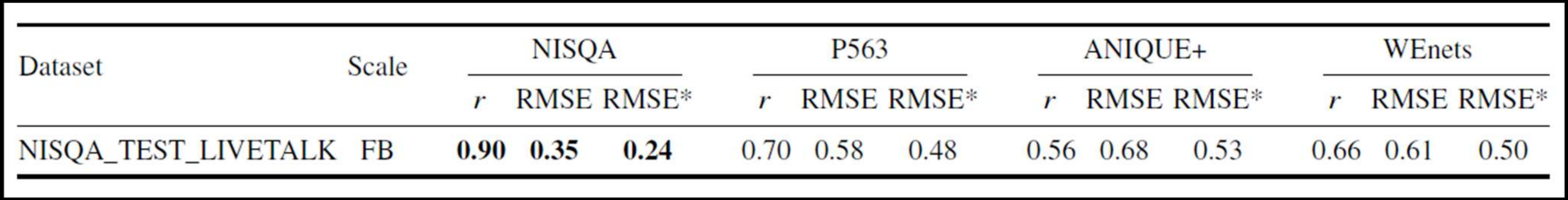
ML Model NISQA

Standard-ization

Ongoing Work

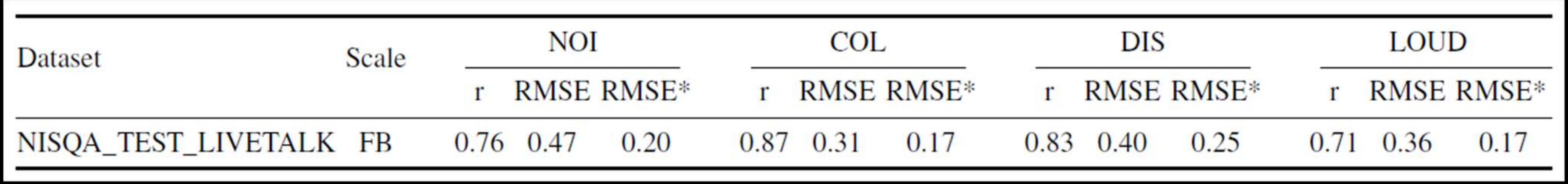
Highlights

Test set results for overall quality



Dataset	Scale	NISQA			P563			ANIQUE+			WEnets		
		r	RMSE	RMSE*	r	RMSE	RMSE*	r	RMSE	RMSE*	r	RMSE	RMSE*
NISQA_TEST_LIVETALK	FB	**0.90**	**0.35**	**0.24**	0.70	0.58	0.48	0.56	0.68	0.53	0.66	0.61	0.50

Test set results for speech quality dimensions



Dataset	Scale	NOI			COL			DIS			LOUD		
		r	RMSE	RMSE*	r	RMSE	RMSE*	r	RMSE	RMSE*	r	RMSE	RMSE*
NISQA_TEST_LIVETALK	FB	0.76	0.47	0.20	0.87	0.31	0.17	0.83	0.40	0.25	0.71	0.36	0.17

The model is evaluated on the test set that was not used during the training or selection of the model, and that contains live-talking conditions, which are independent of the conditions and talkers contained in the other datasets.

G. Mittag, B. Naderi, A. Chehadi, and S. Möller “NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets,” in Proc. Interspeech 2021, 2021.

Speech Quality

Quality Prediction

ML Model NISQA

Standard-ization

Ongoing Work

Highlights

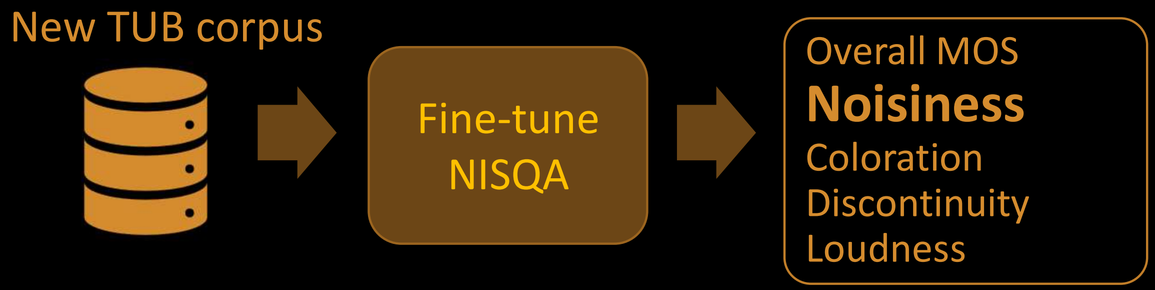
Standardization

ITU-T Study Group 12

Question 9

Single-ended perceptual approaches for multi-dimensional analysis – P.SAMD

NISQA has been submitted here for standardization and is the only P-SAMD candidate model.



INTERSPEECH 2022
September 18 - 22 • Incheon Korea

ConferencingSpeech 2022 Challenge

(MOS only)

Multiple new datasets provided

Simplified NISQA Baselines

New approaches

New technique:

Large pre-trained language models for feature extraction

(Wave2vec2, XLS-R)

Speech Quality

Quality Prediction

ML Model NISQA

Standard-ization

Ongoing Work

Highlights

NISQA model

[Non-intrusive speech quality assessment]
by Gabriel Mittag, Ph.D.

Open sourced

Python - PyTorch

Trained weights

Various versions

Several datasets

https://github.com/gabrielmittag/NISQA.git

G. Mittag, B. Naderi, A. Chehadi, and S. Möller “NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets,” in Proc. Interspeech 2021, 2021.