Microsoft Research

Each year Microsoft Research hosts hundreds of influential speakers from around the world including leading scientists, renowned experts in technology, book authors, and leading academics, and makes videos of these lectures freely available.

AUDIO QUALITY ASSESSMENT WITH DEEP NEURAL NETWORKS

Anderson Raymundo Avila, INRS-EMT

Mentor: Ivan Tashev, Microsoft Research

OUTLINE

1. Introduction

2. Dataset

3. Proposed Approaches

4. Results and Conclusion

AUDIO QUALITY ASSESSMENT

Prediction of audio quality level

Instrumental measure that can predict the audio quality level as perceived by humans

QUALITY OF EXPERIENCE (QoE)

- Can you hear me?
- Speakers interrupting each other
- Distortion in the voice
- Unnaturalness

NETWORK IMPAIRMENTS

- Latency
- Jitter
- Packet loss

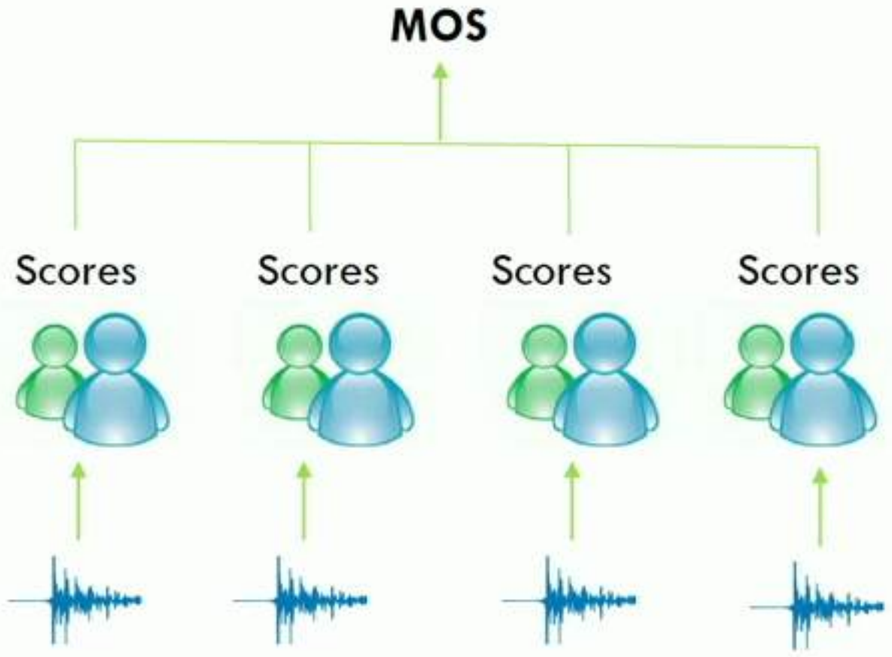
- Additive noise
- Convolutive noise
- Clipping
- Unwanted artifacts introduced by enhancement algorithms

SPEECH DISTORTION

- Latency
- Jitter
- Packet loss

- Additive noise
- Reverberation
- Clipping
- Artifacts introduced by enhancement algorithms

SUBJECTIVE LISTENING TESTS



- **Most reliable**
- Time-consuming
- Laborious
- Expensive
- Cannot be done in real-time
- ITU-T P.800

INSTRUMENTAL MEASURES

- Signal-based, parametric-based or hybrid
- Single and double-ended (PESQ), elsewhere denoted as non-intrusive and intrusive
- Computational proxy of listening tests

SIGNAL-BASED MODELS

Signal-based Models	Recommended For	Limitations
PESQ	Input levels, transmission channel errors, bit-rates, transcodings, noise at sending side, time-varying delay, waveform codecs, celp codecs, other codecs	Listening levels, loudness, hybrid codecs, time-warping, noise reduction, echo cancellation
WB-PESQ	As above, but including WB transmission	As above, WB transmission, hybrid decods such as AMR-WB, G.729 and EVRC-WB
POLQA	Same as PESQ above, but with SWB transmission VQE algorithms, short time-warping, hybrid speech codecs	Strong time-warping distortions
P.563	Same as PESQ and for short and long-term time warping of the speech signal	Strong time-warping distortions, EVRC codecs
ANIQUE+	Same as P.563	Same as P.563

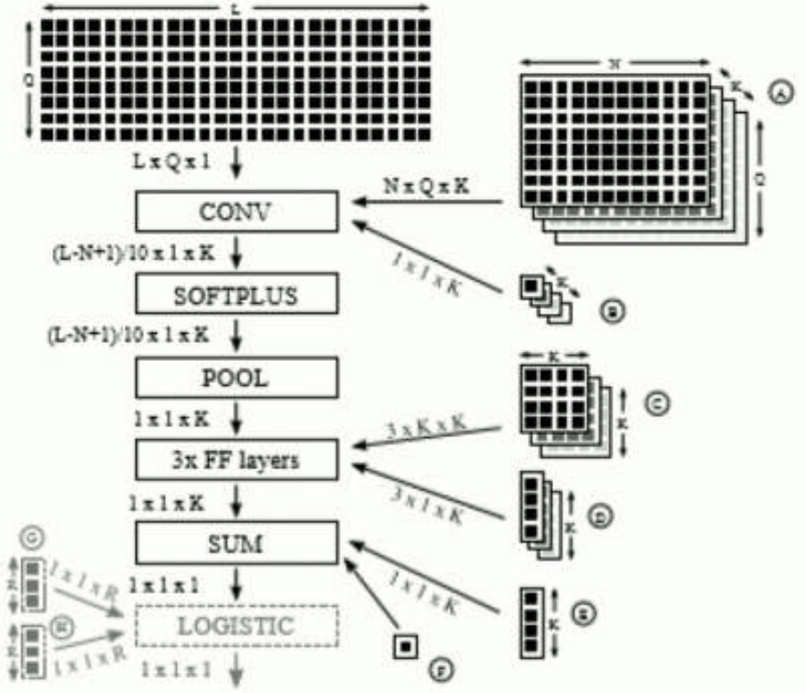
PROJECT MOTIVATION

- PESQ is widely used for audio quality assessment:
  - It was trained only to the distortions introduced by the speech compression
  - It is intrusive, i.e. limits us to synthetic datasets
- We need new algorithm for objective audio quality measurement:
  - Trained on noise, reverberation, and audio pipeline distortions
  - Non-intrusive, so we can monitor the call quality in real time

LITERATURE REVIEW

- Soni, Meet H., and Hemant A. Patil. "Novel deep auto-encoder features for non-intrusive speech quality assessment." *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016.
- Spille, Constantin, et al. "Predicting speech intelligibility with deep neural networks." *Computer Speech & Language* 48 (2018).
- Huber, R., Krüger M. and Bernd T. M. "Single-ended prediction of listening effort using deep neural networks." *Hearing research* 359 (2018).
- Andersen, A. H., et al. "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018).
- Fu, Szu-Wei, et al. "Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model based on BLSTM". In *Proceedings INTERSPEECH 2018, International Speech Communication Association, Hyderabad, India*.
- Ooster, J., Huber, R. and Bernd, T. M. "Prediction of Perceived Speech Quality Using Deep Machine Learning". In *Proceedings INTERSPEECH 2018, International Speech Communication Association, Hyderabad, India*.

LITERATURE REVIEW

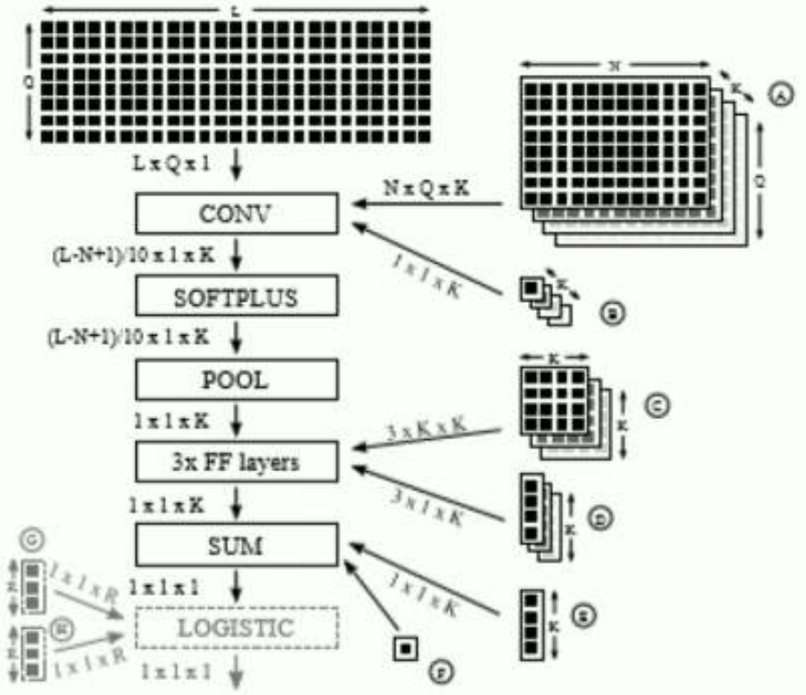


Andersen, A. H., et al. "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018).

LITERATURE REVIEW

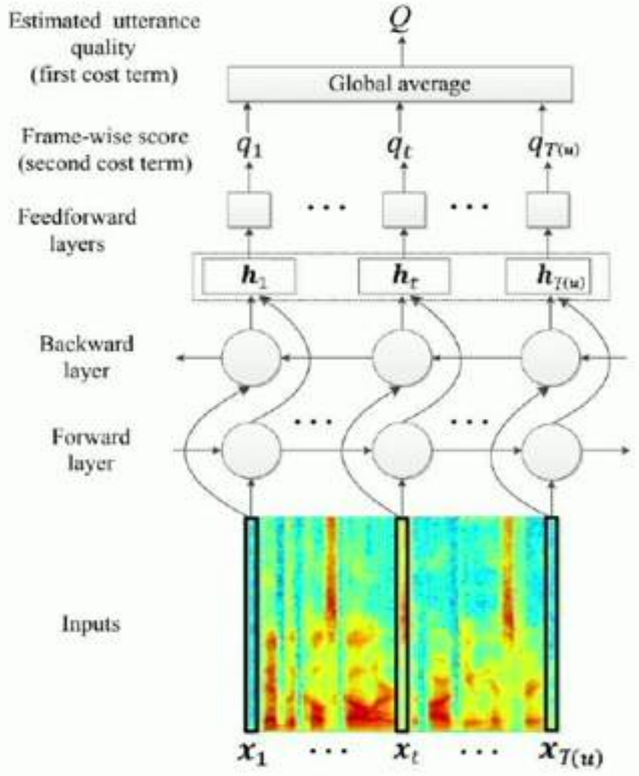
- Soni, Meet H., and Hemant A. Patil. "Novel deep auto-encoder features for non-intrusive speech quality assessment." *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016.
- Spille, Constantin, et al. "Predicting speech intelligibility with deep neural networks." *Computer Speech & Language* 48 (2018).
- Huber, R., Krüger M. and Bernd T. M. "Single-ended prediction of listening effort using deep neural networks." *Hearing research* 359 (2018).
- Andersen, A. H., et al. "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018).
- Fu, Szu-Wei, et al. "Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model based on BLSTM". In *Proceedings INTERSPEECH 2018, International Speech Communication Association, Hyderabad, India*.
- Ooster, J., Huber, R. and Bernd, T. M. "Prediction of Perceived Speech Quality Using Deep Machine Learning". In *Proceedings INTERSPEECH 2018, International Speech Communication Association, Hyderabad, India*.

LITERATURE REVIEW



Andersen, A. H., et al. "Nonintrusive Speech Intelligibility Prediction Using Convolutional Neural Networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018).

LITERATURE REVIEW



$$O = \frac{1}{S}\sum_{s=1}^{S}\left[(\hat{Q}_s - Q_s)^2 + \alpha(\hat{Q}_s)\sum_{t=1}^{T(u_s)}(\hat{Q}_s - q_{s,t})^2\right]$$

Fu, Szu-Wei, et al. "Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model based on BLSTM". In *Proceedings INTERSPEECH 2018, International Speech Communication Association, Hyderabad, India.*

OUTLINE

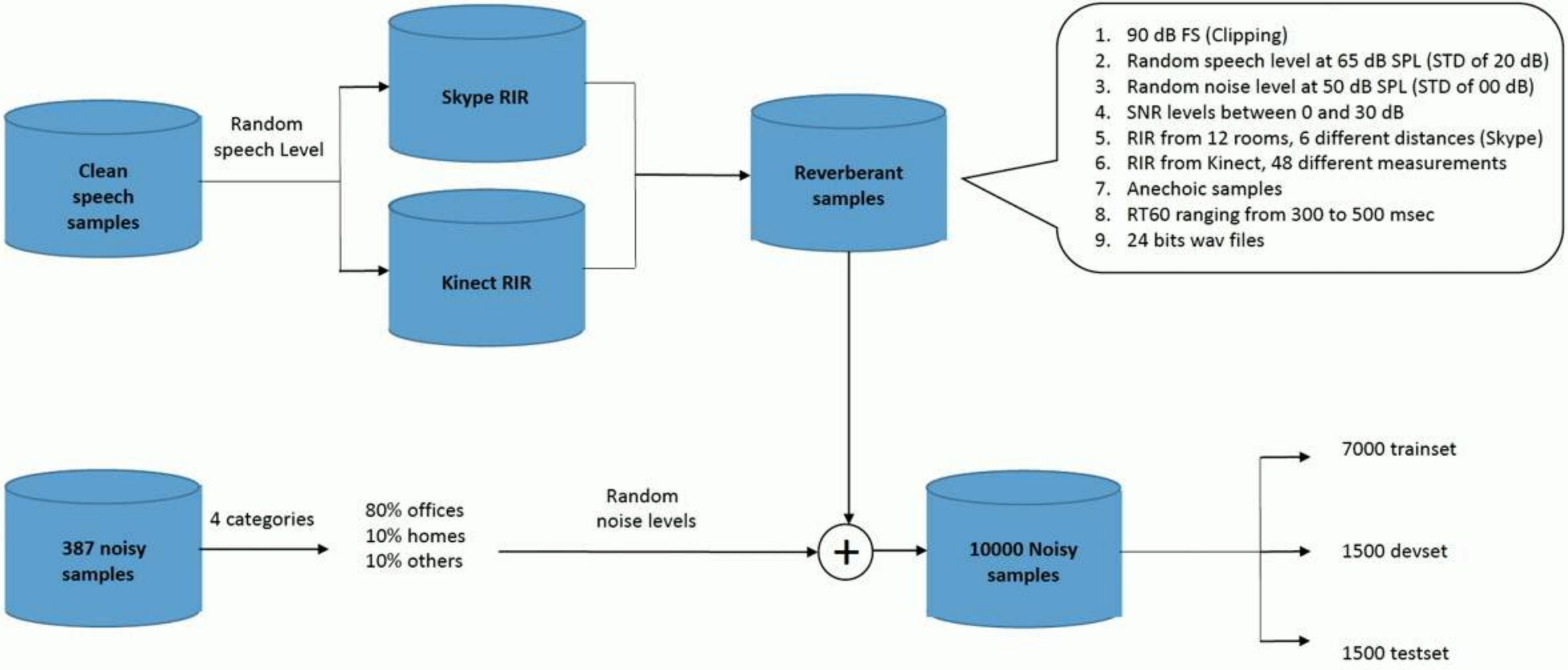
1. Introduction

2. Dataset

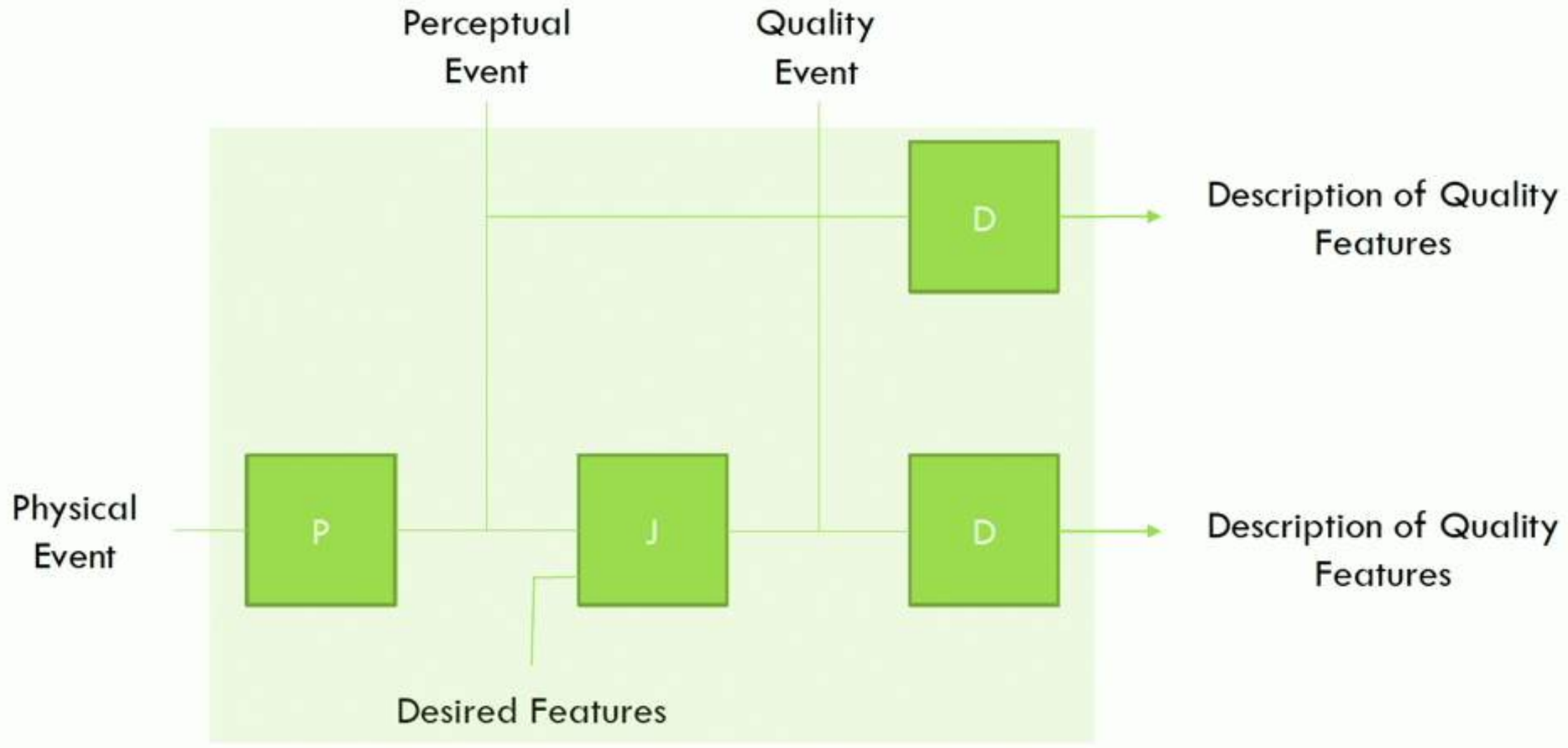
3. Proposed Approaches

4. Results and Conclusion

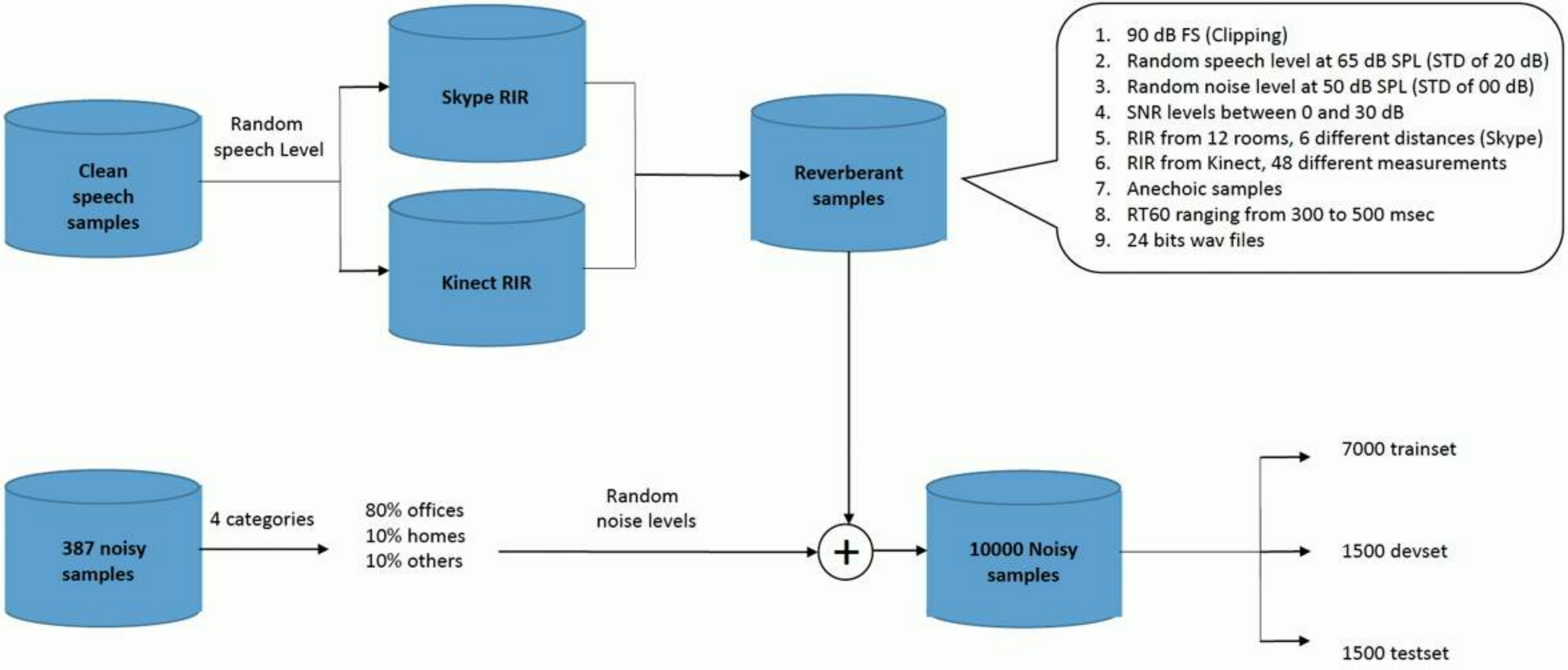
DATA GENERATION



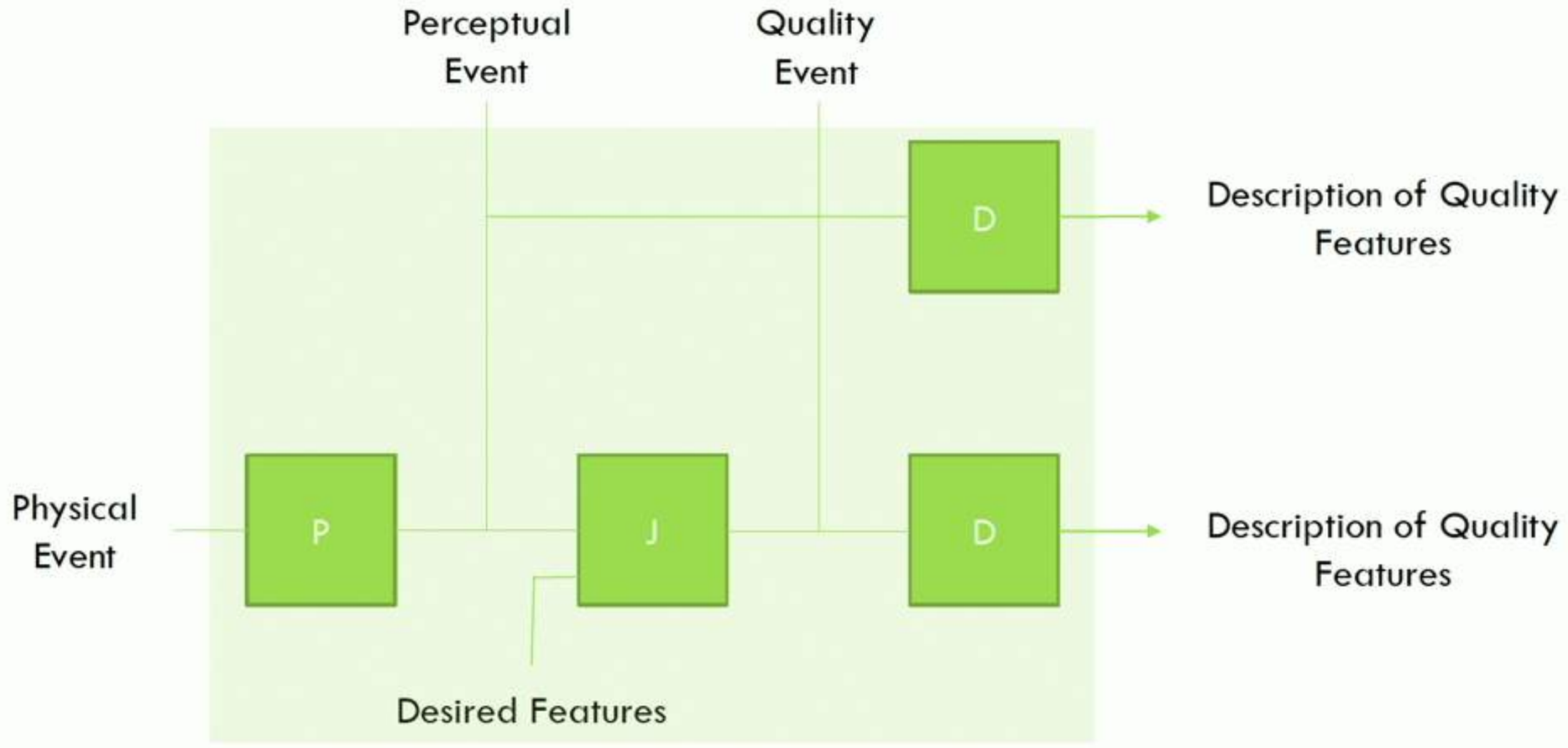
LISTENING QUALITY TEST



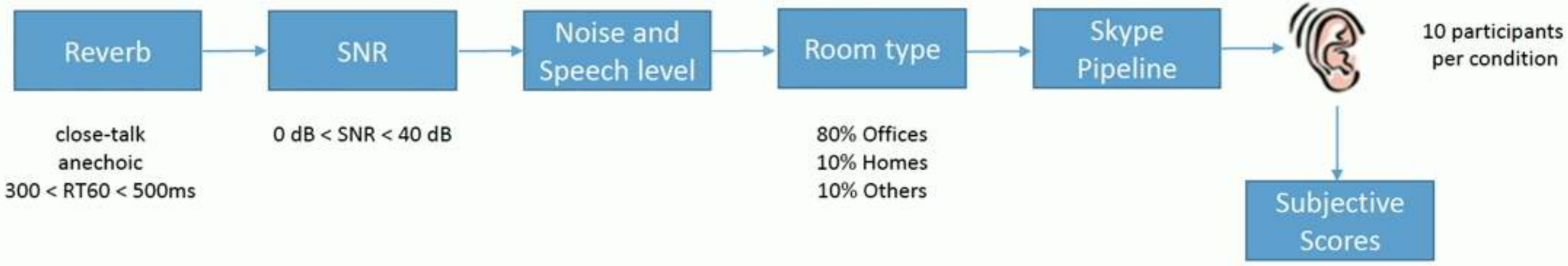
DATA GENERATION



LISTENING QUALITY TEST

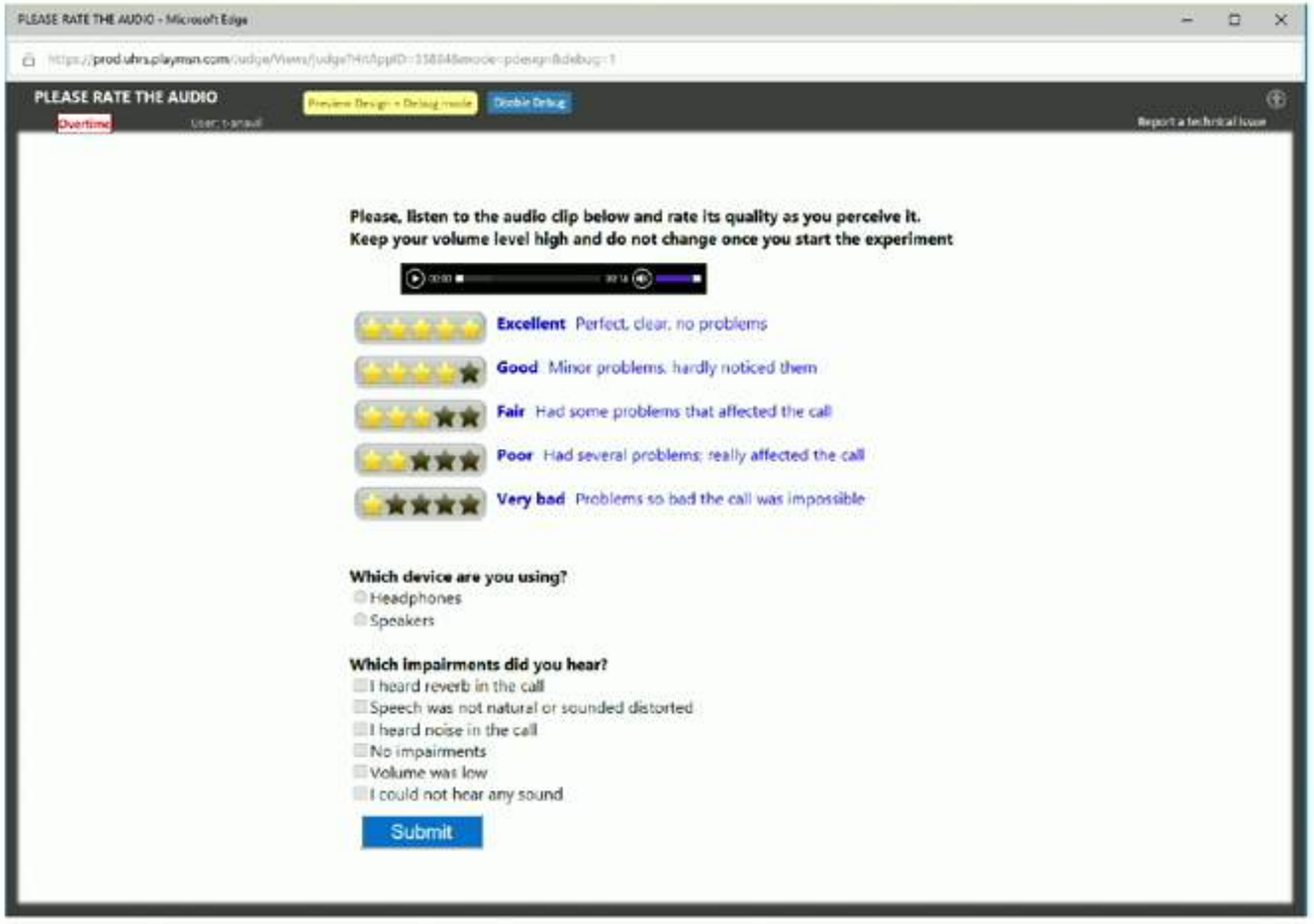


LISTENING QUALITY TEST

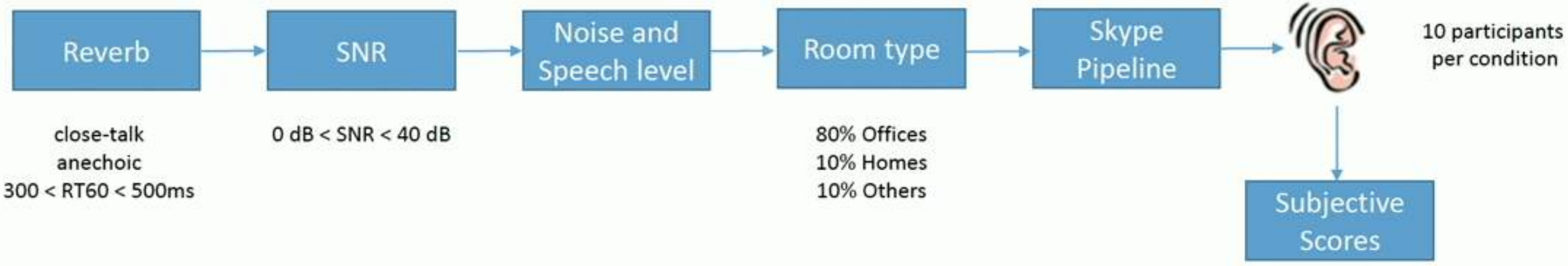


Labeling has done using crowdsourcing.
10,000 files x 10 judges = 100,000 labels.

LISTENING QUALITY TEST

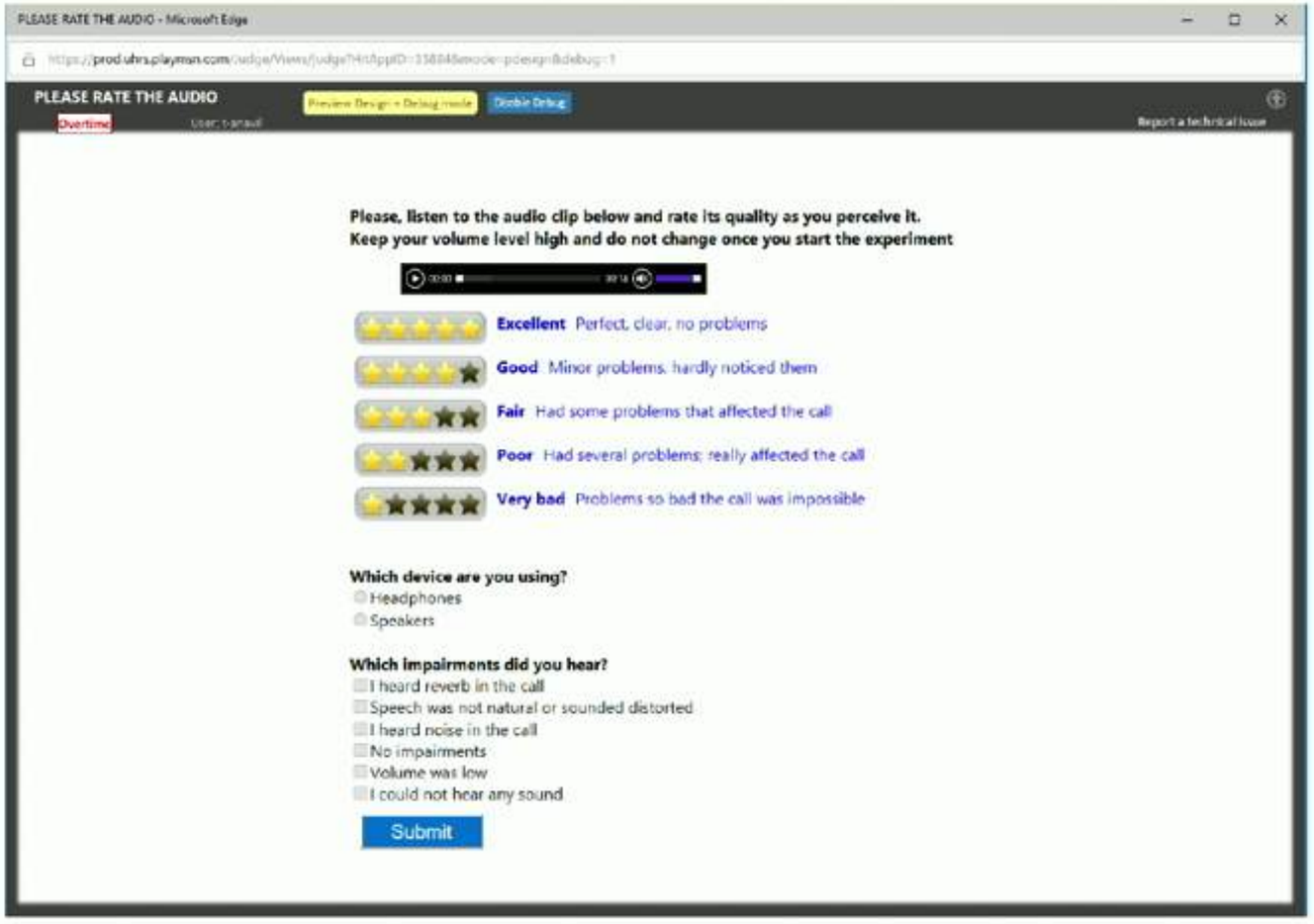


LISTENING QUALITY TEST

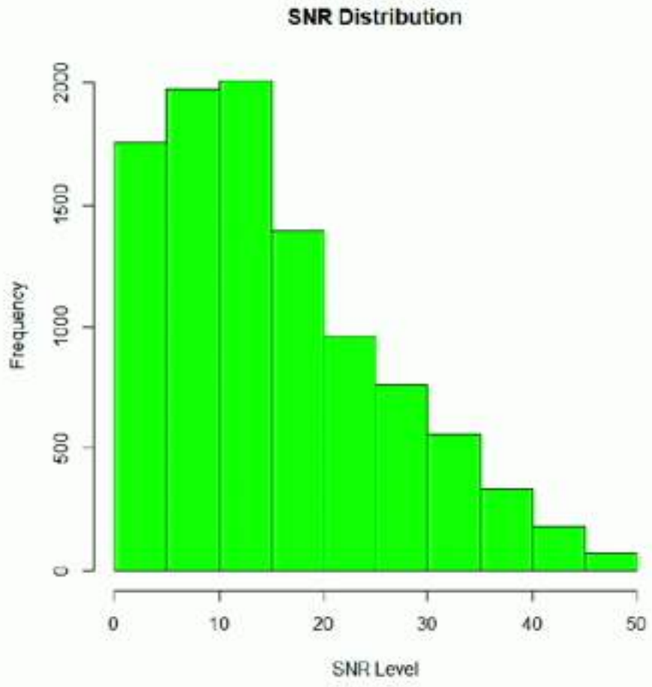


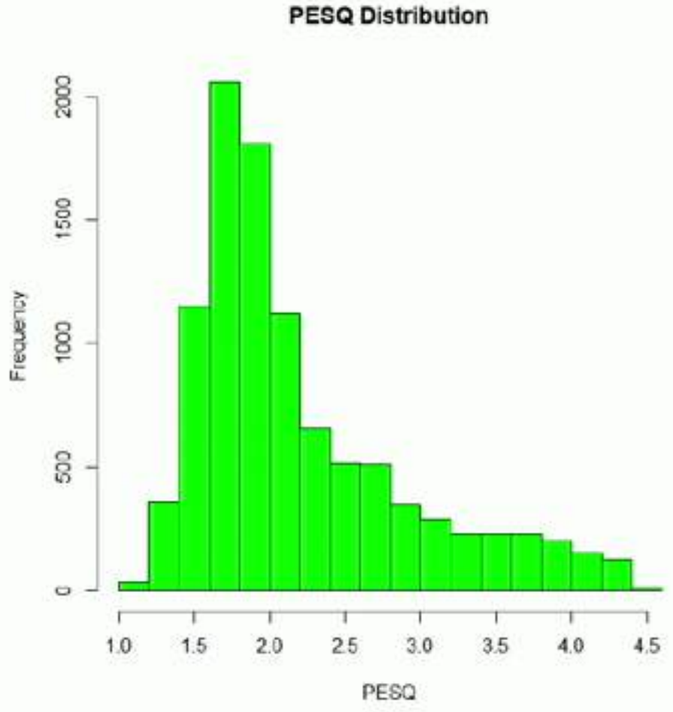
Labeling has done using crowdsourcing.
10,000 files x 10 judges = 100,000 labels.

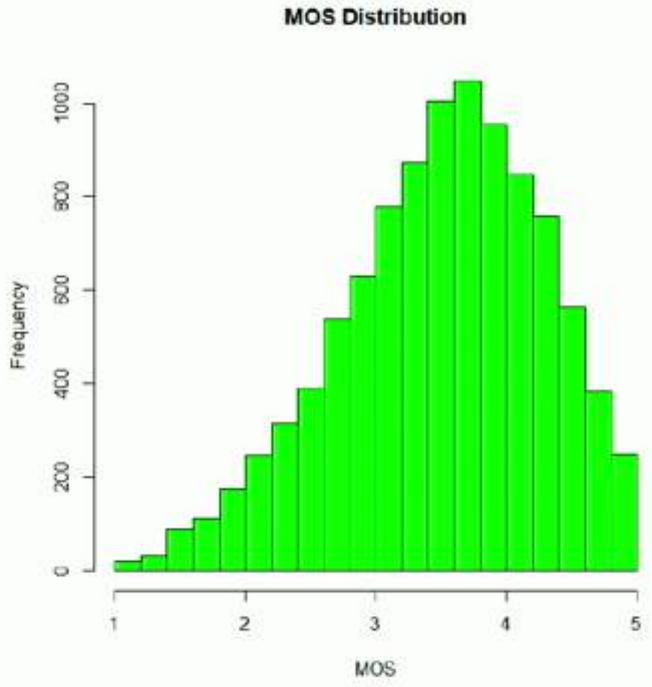
LISTENING QUALITY TEST



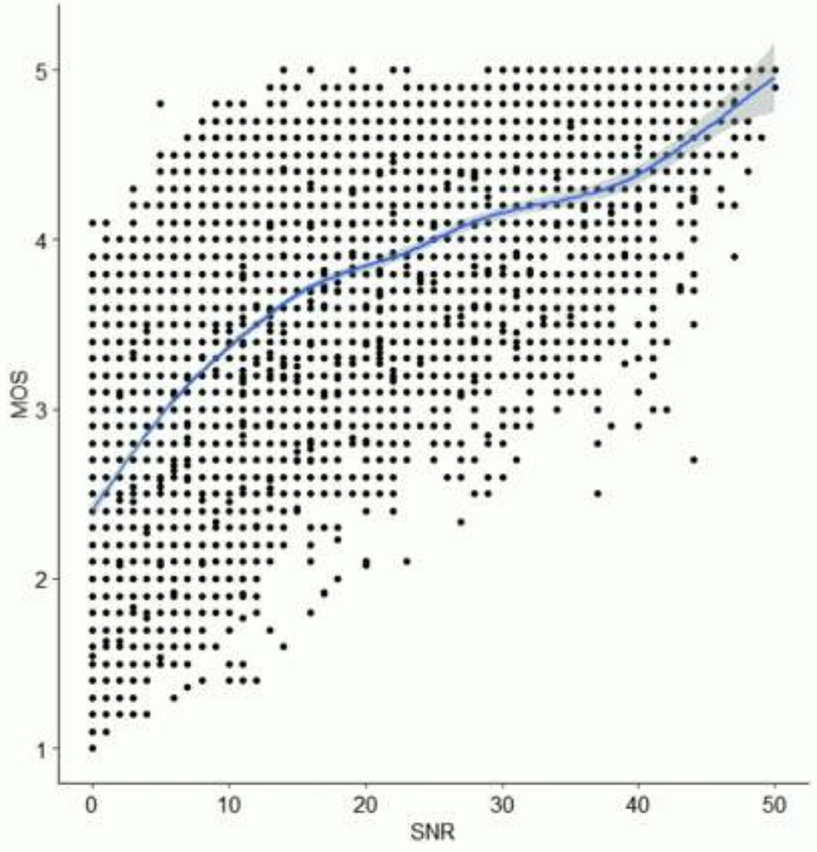
DATA EXPLORATORY ANALYSIS

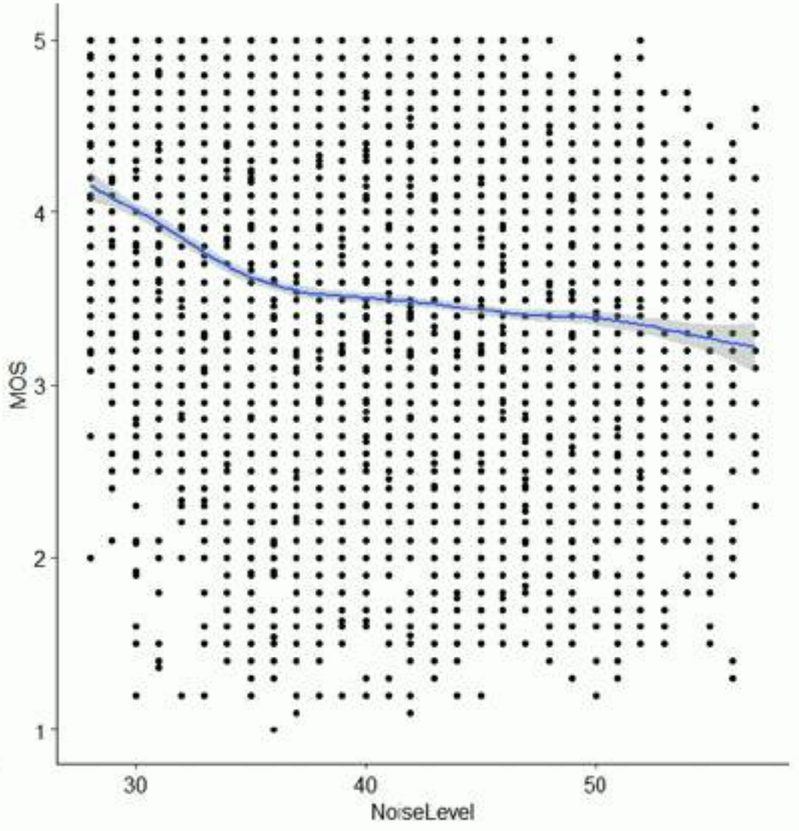


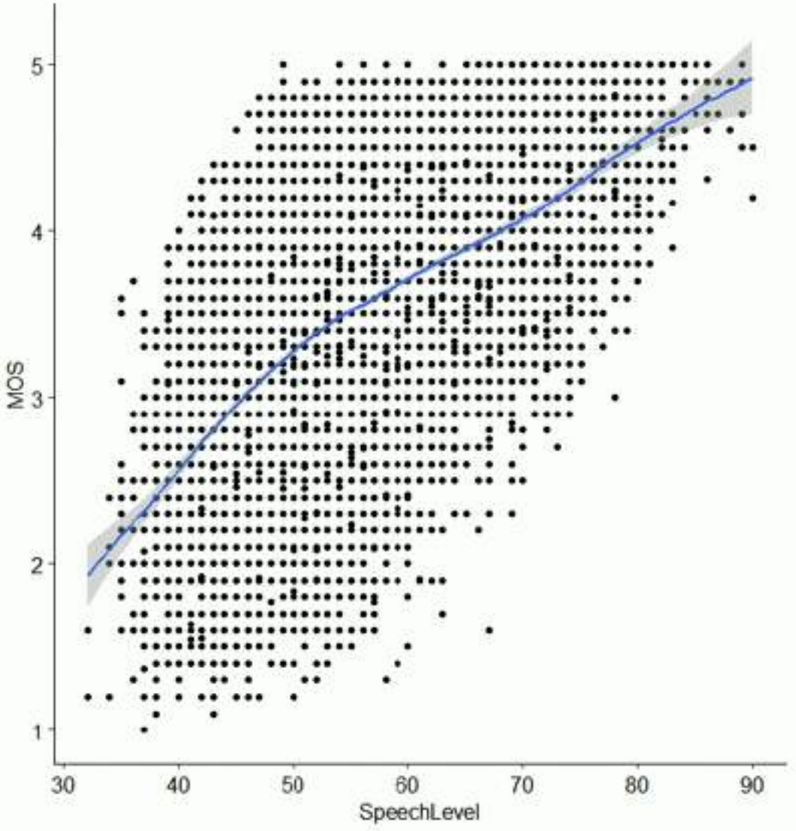




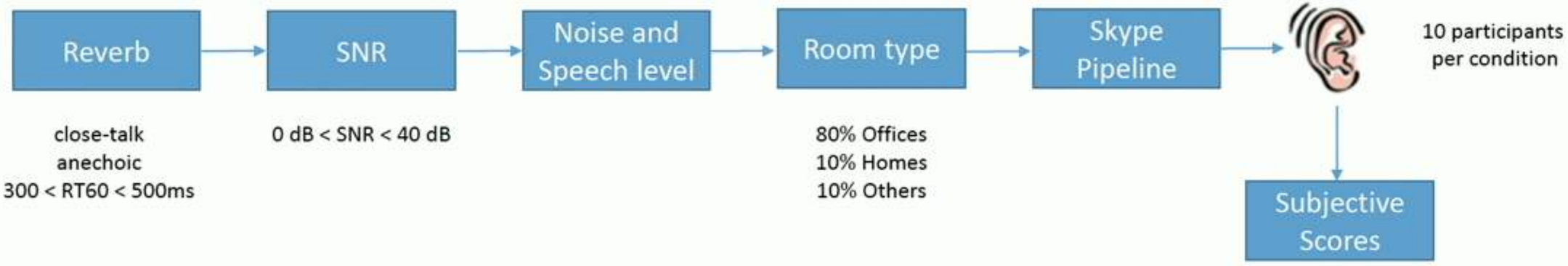
DATA EXPLORATORY ANALYSIS





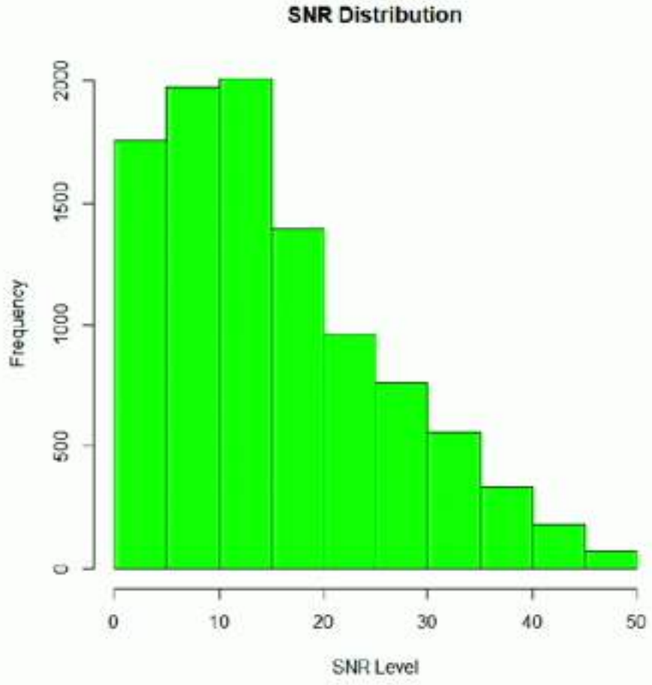


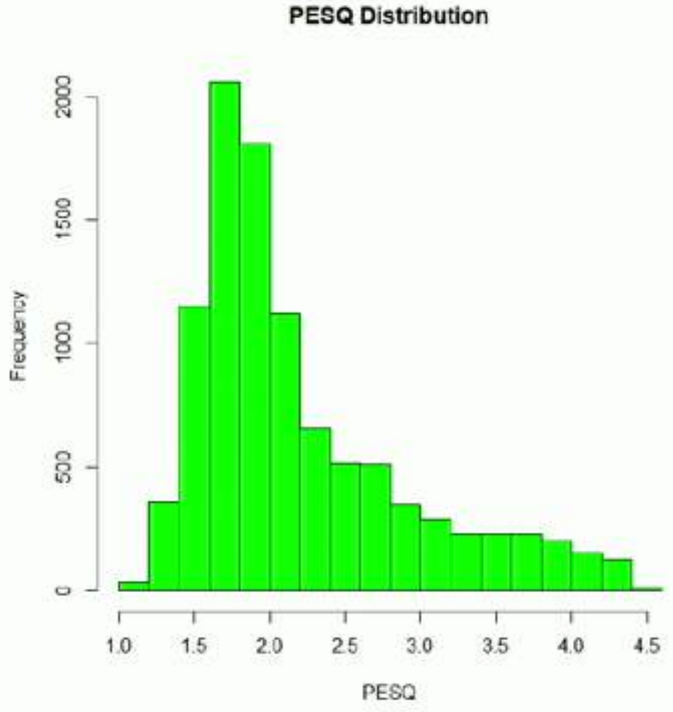
LISTENING QUALITY TEST

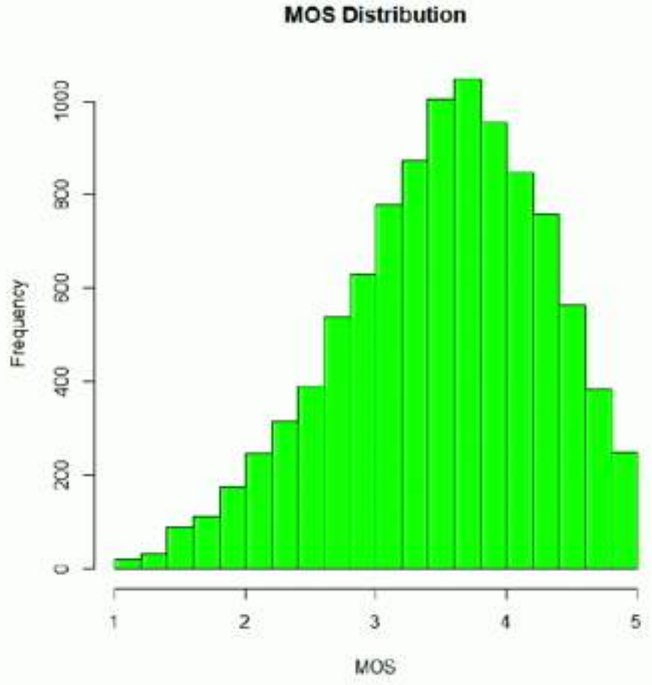


Labeling has done using crowdsourcing.
10,000 files x 10 judges = 100,000 labels.

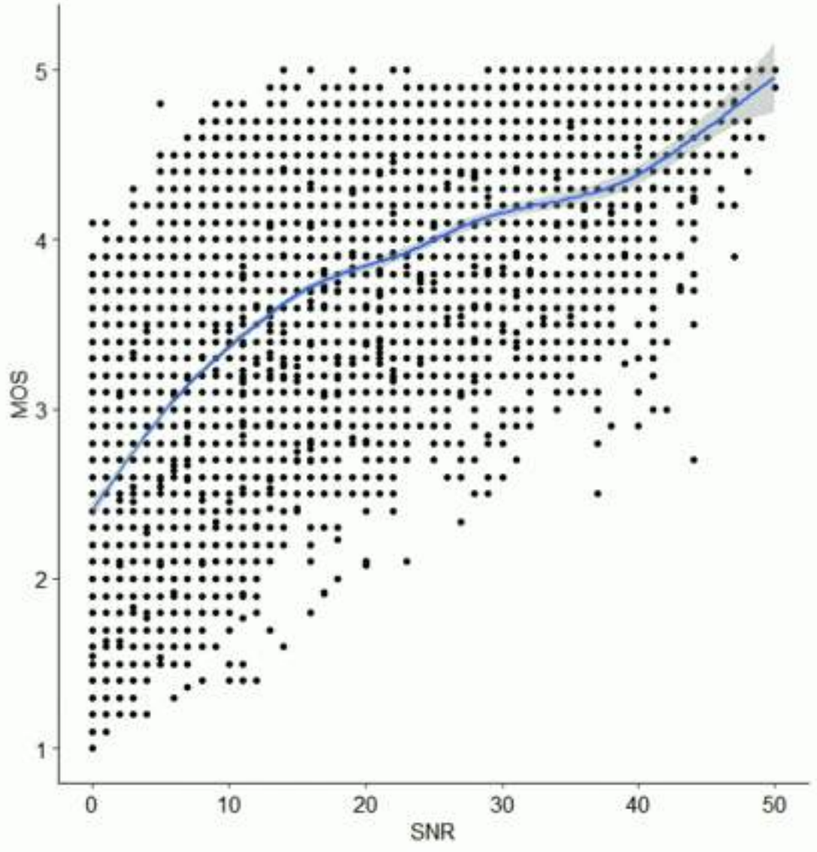
DATA EXPLORATORY ANALYSIS

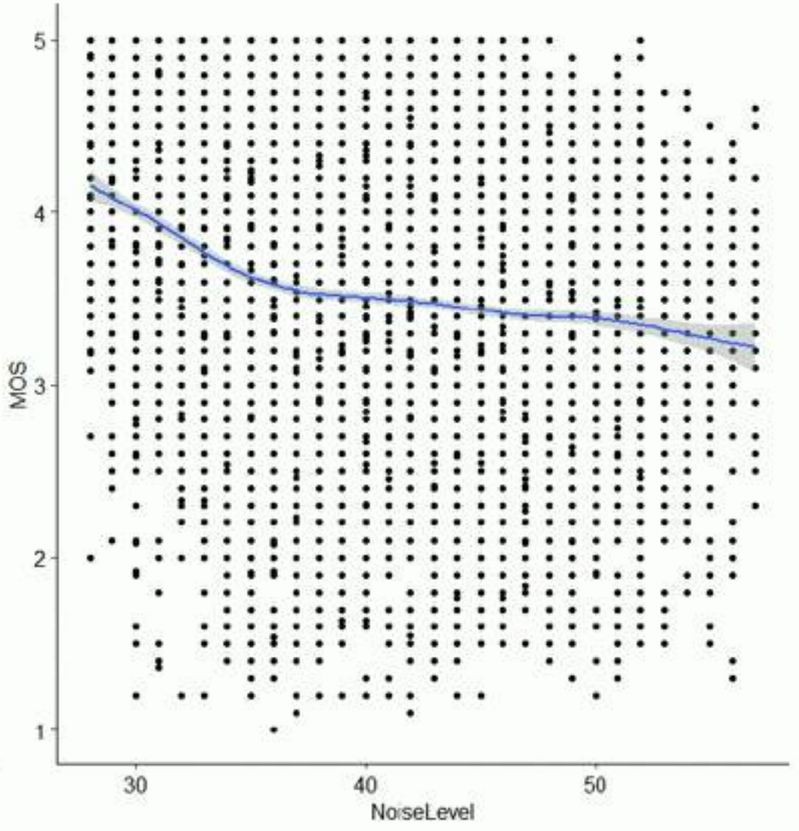


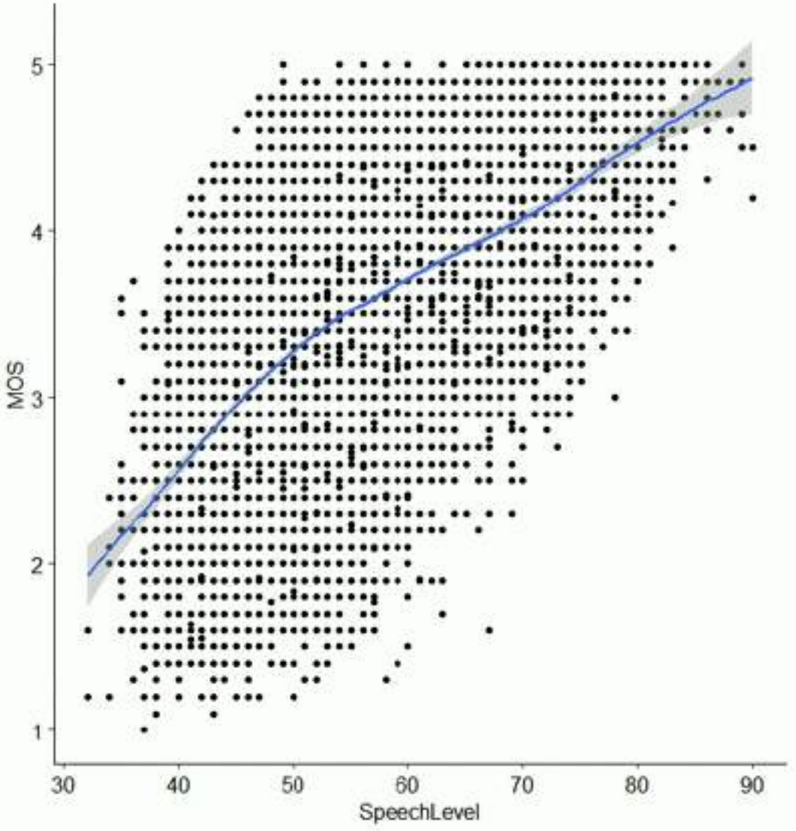




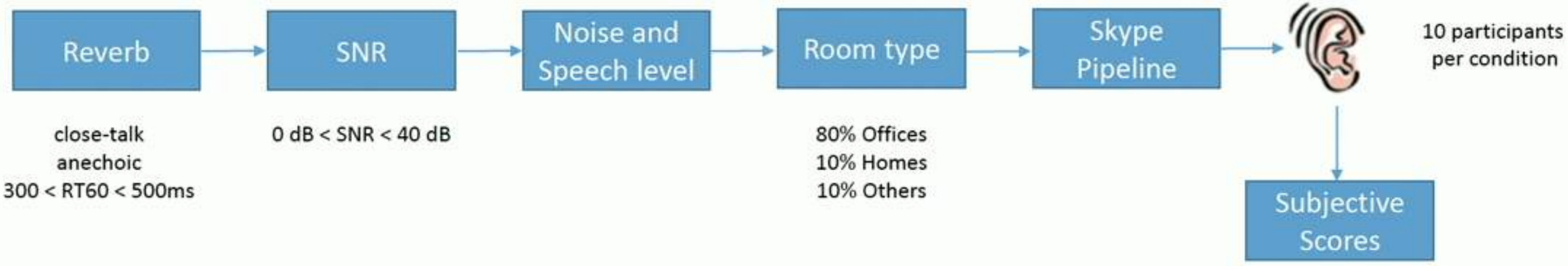
DATA EXPLORATORY ANALYSIS





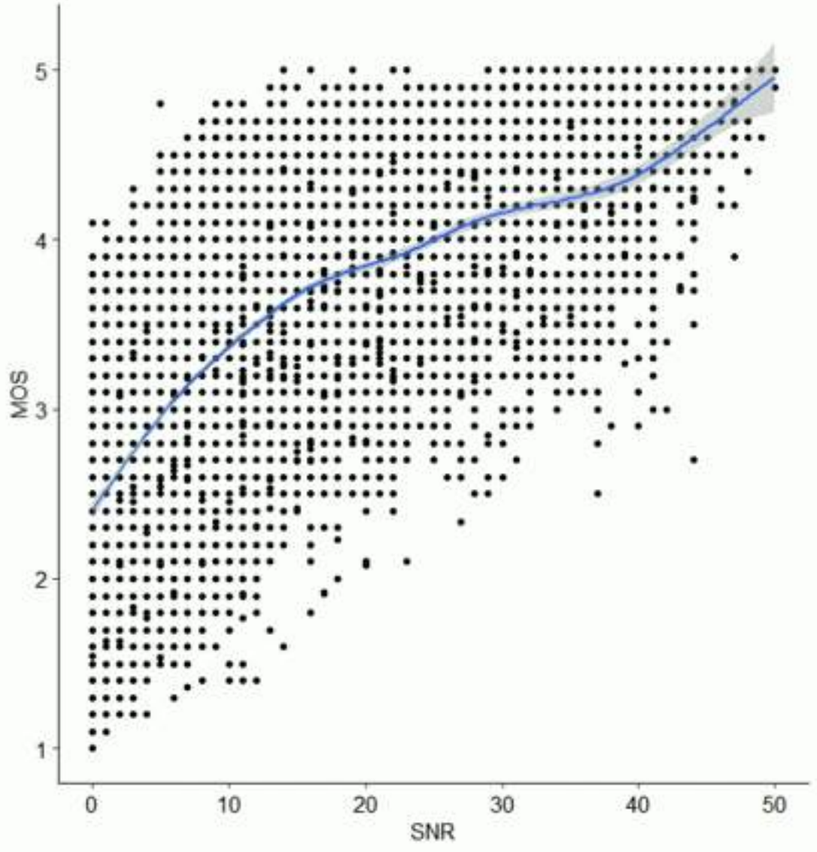


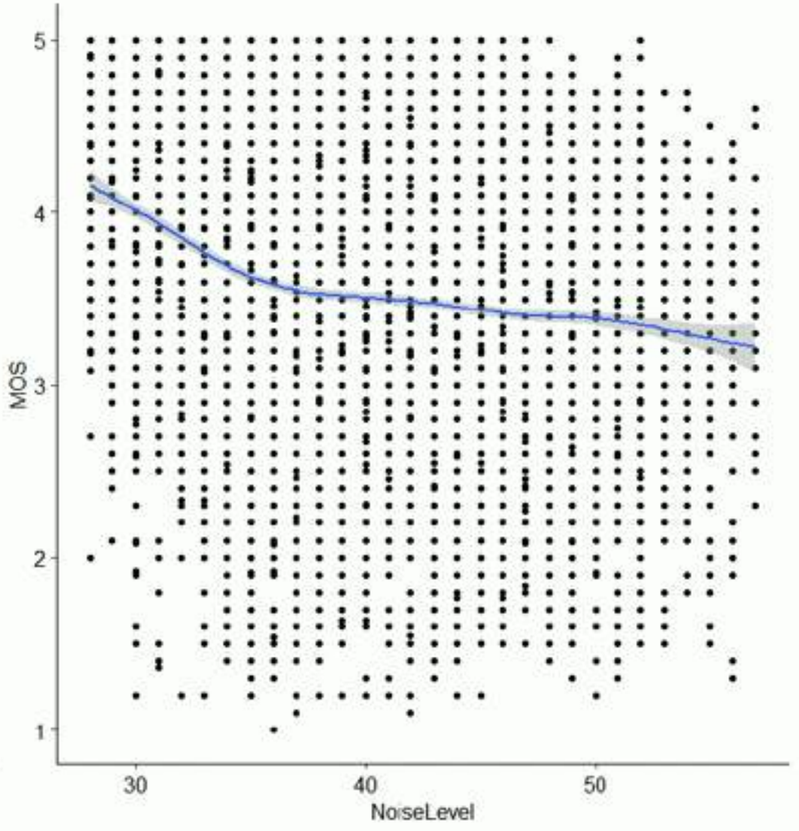
LISTENING QUALITY TEST

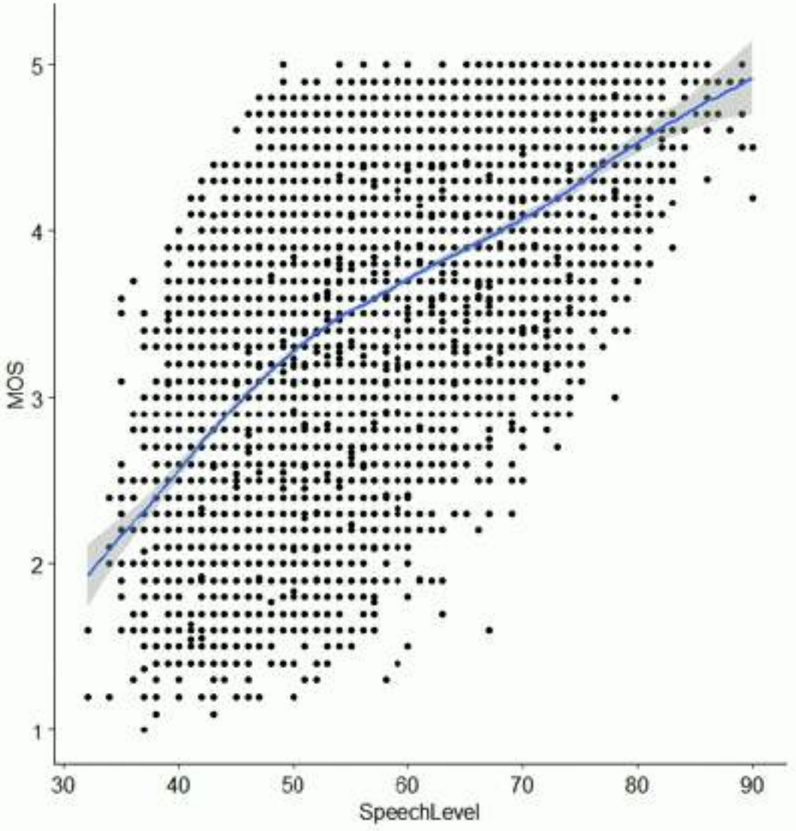


Labeling has done using crowdsourcing.
10,000 files x 10 judges = 100,000 labels.

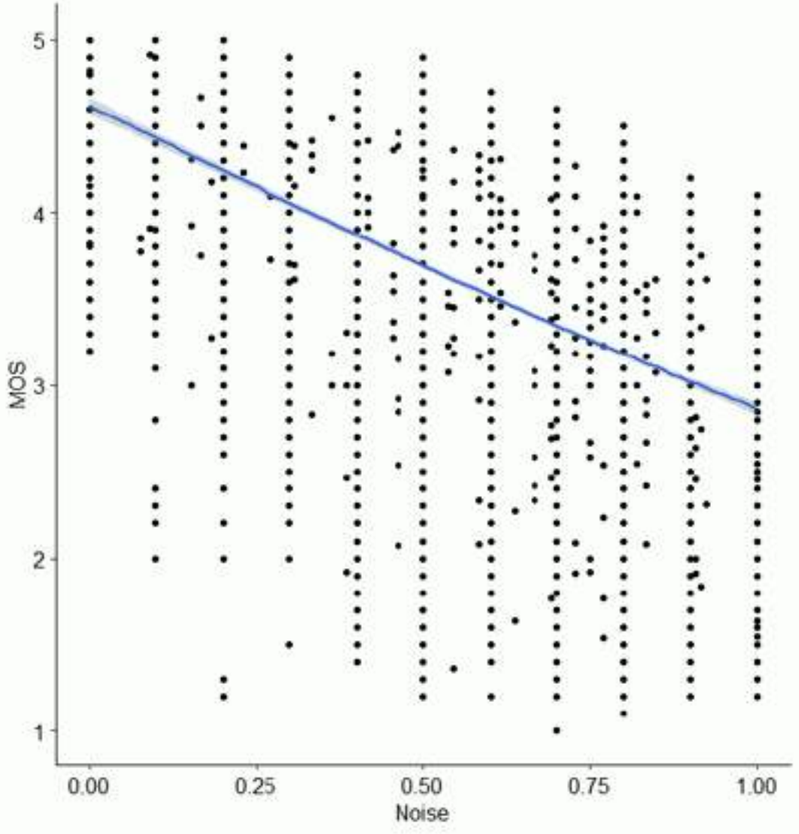
DATA EXPLORATORY ANALYSIS

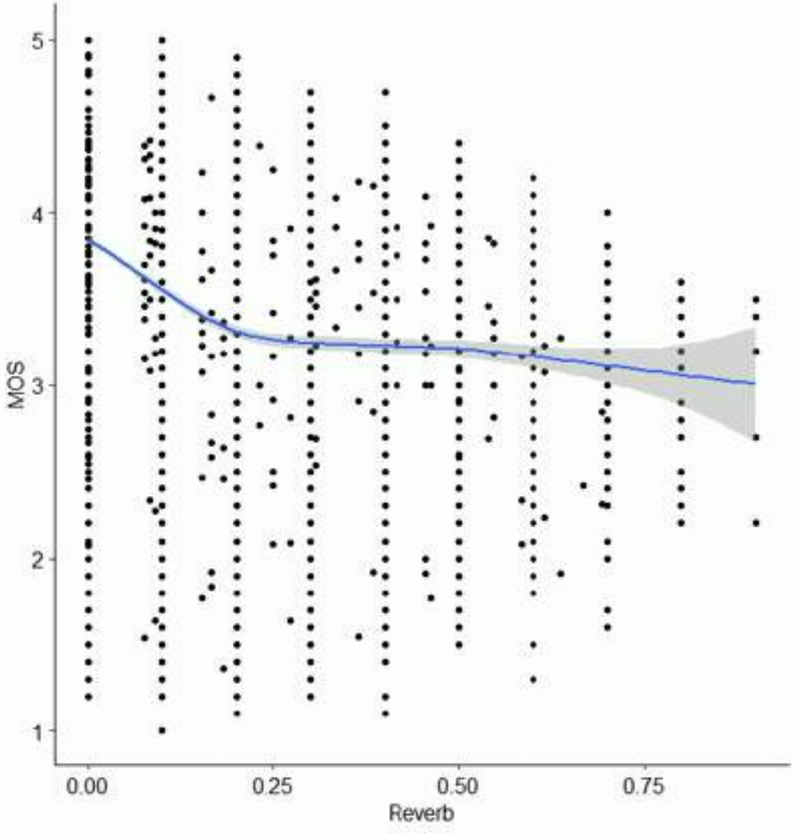


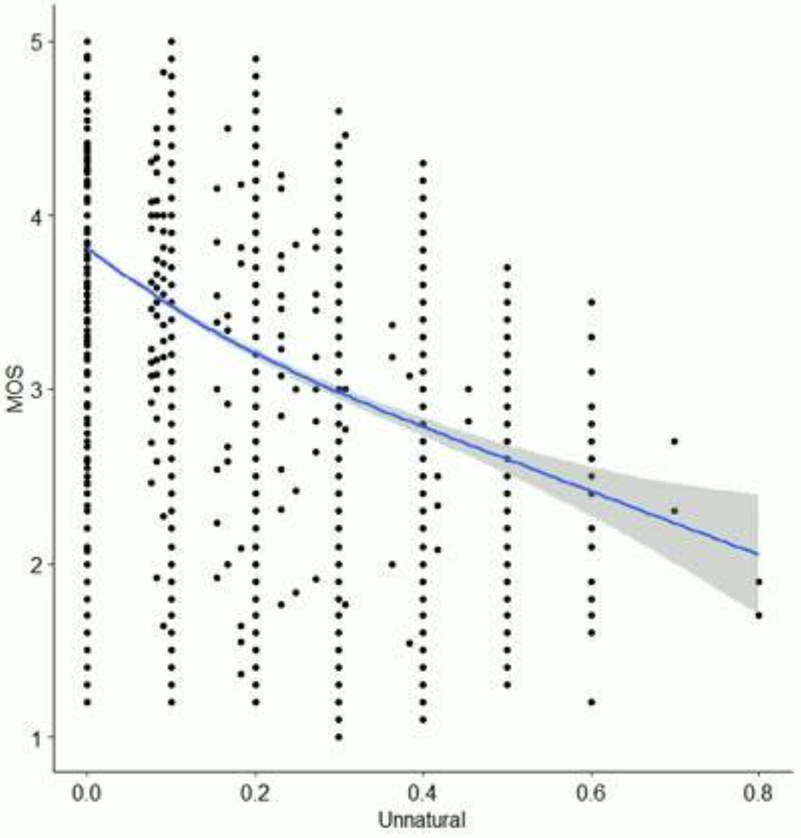




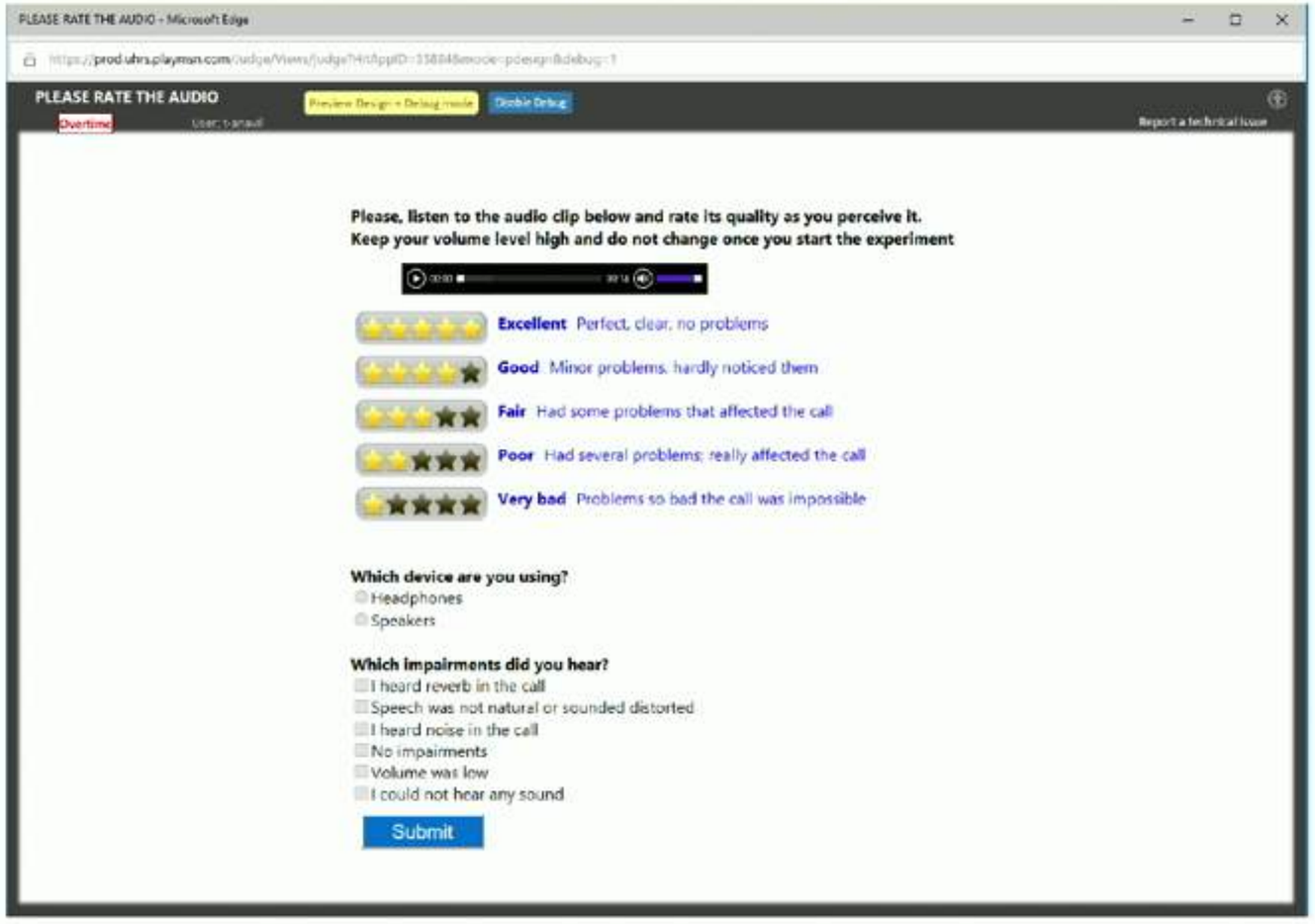
DATA EXPLORATORY ANALYSIS



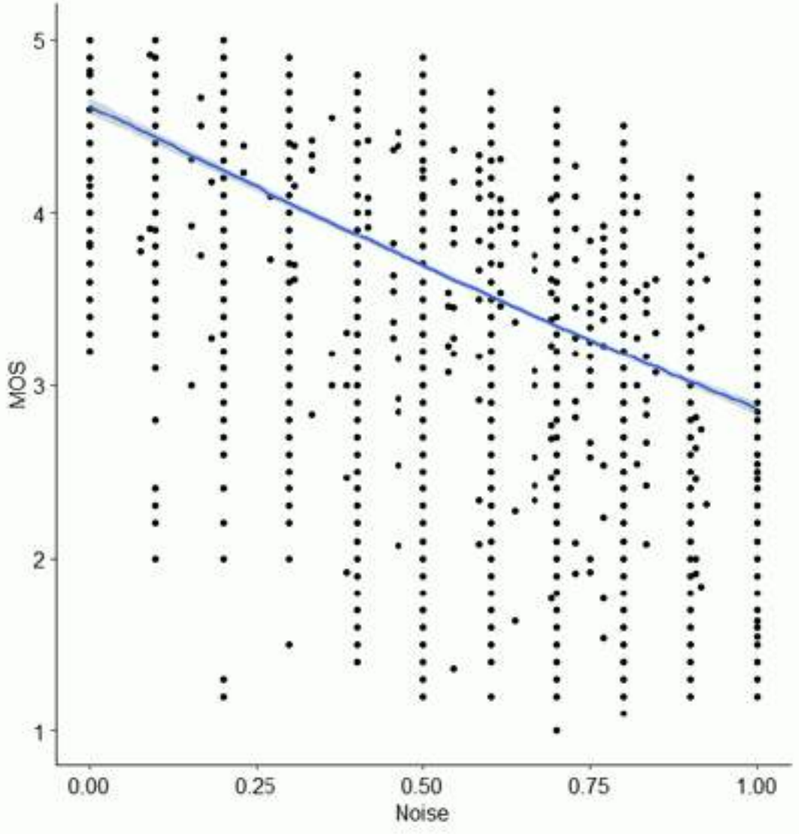


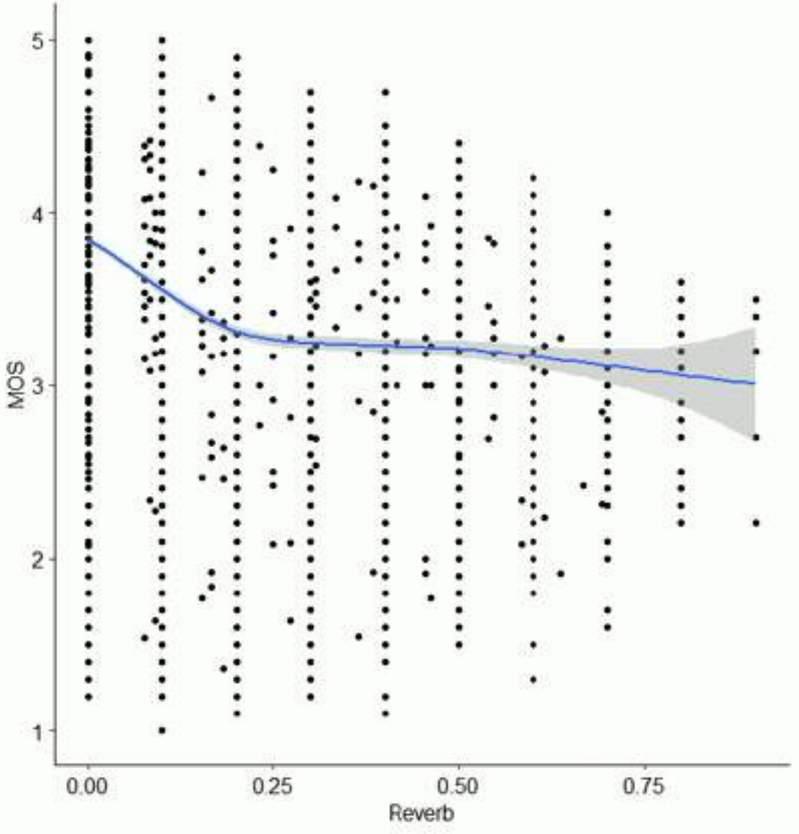


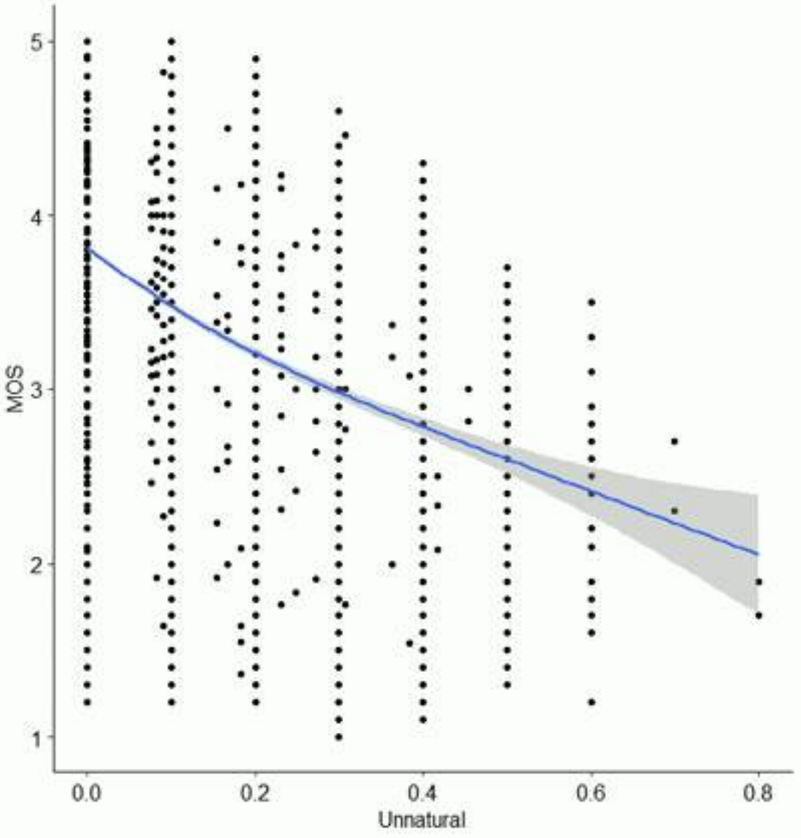
LISTENING QUALITY TEST



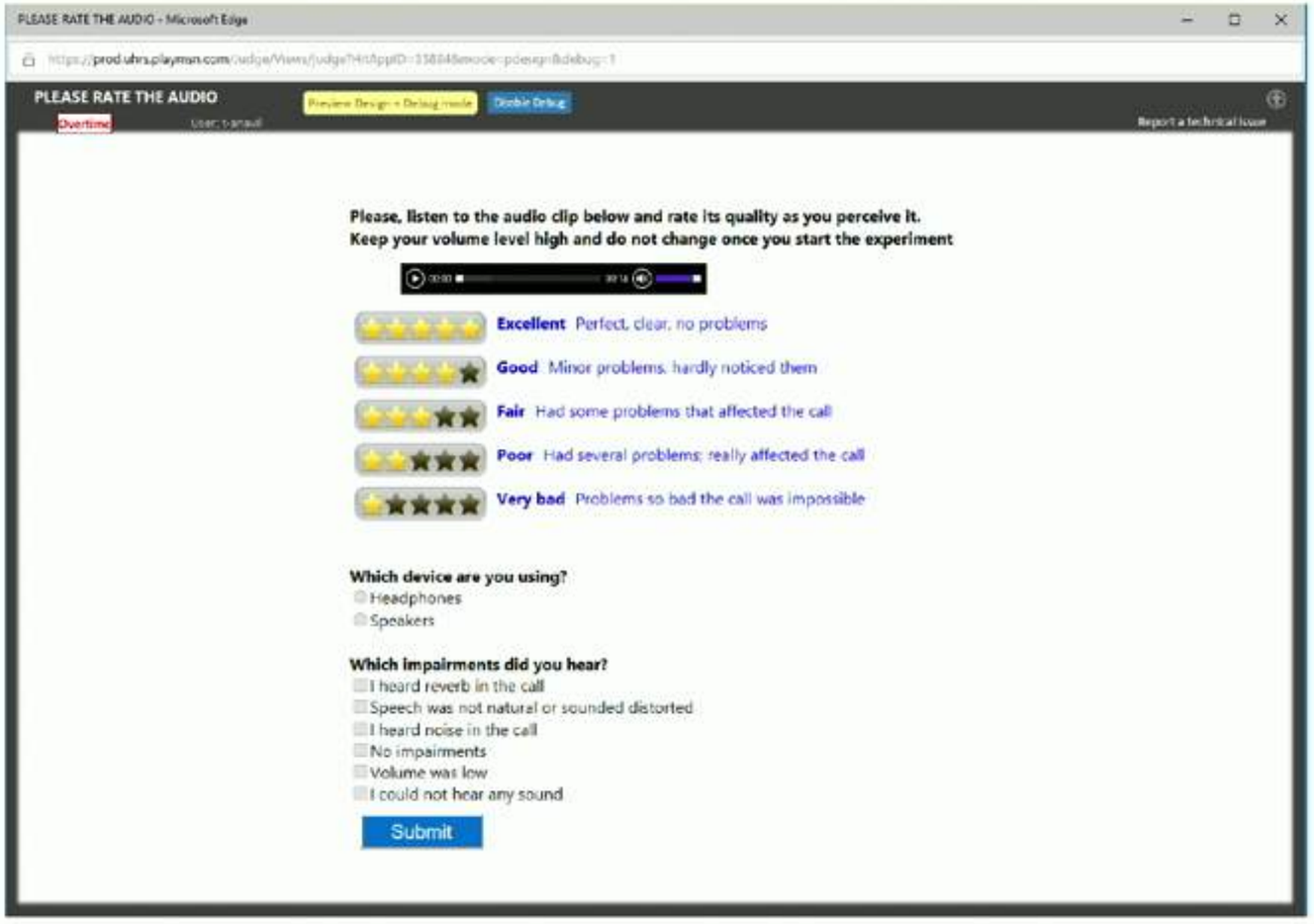
DATA EXPLORATORY ANALYSIS



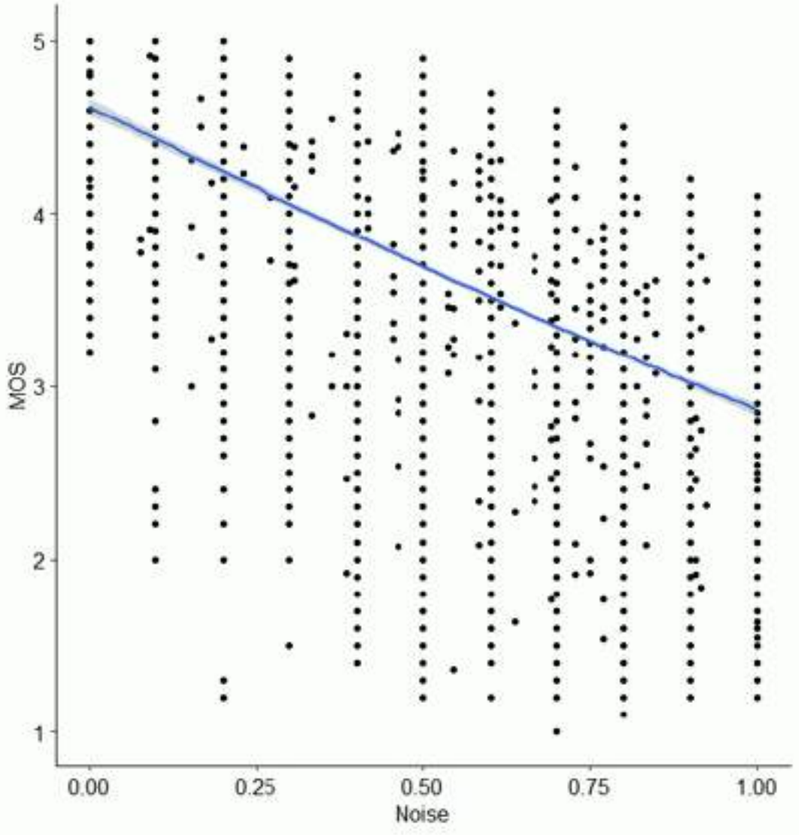


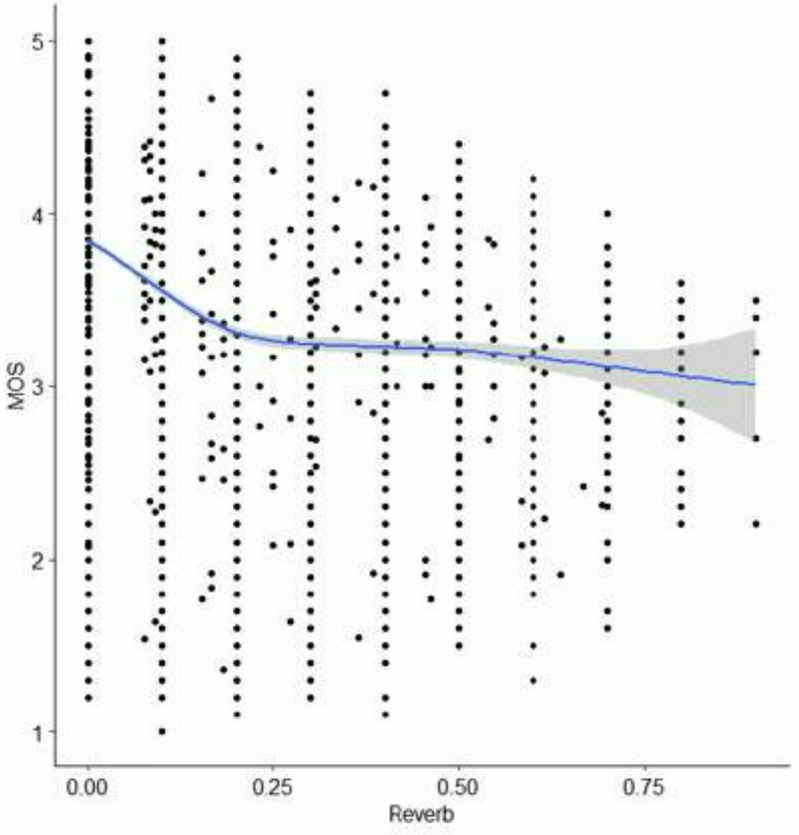


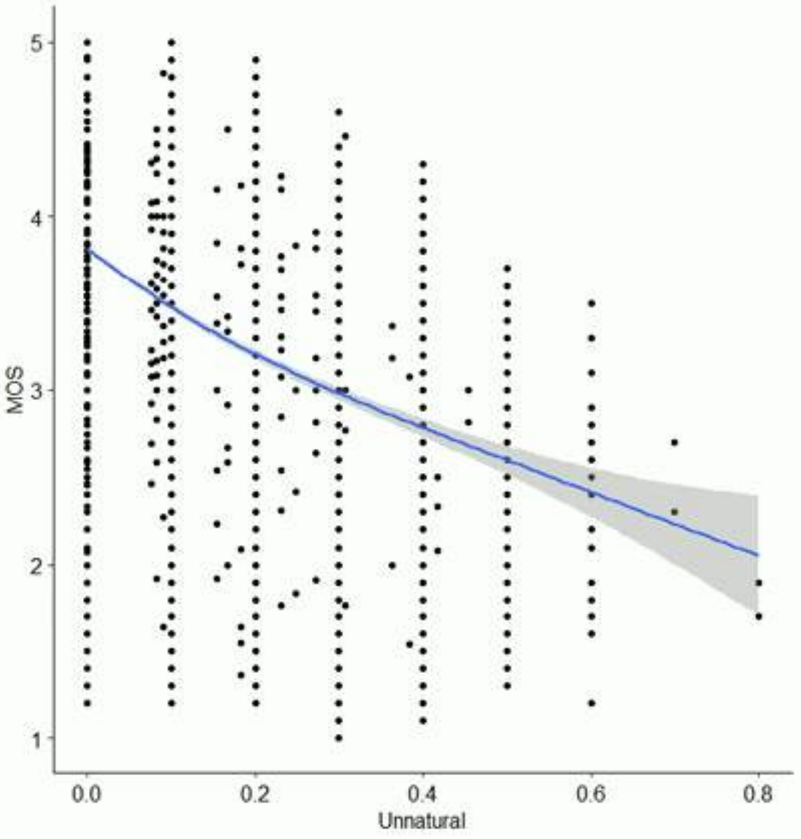
LISTENING QUALITY TEST



DATA EXPLORATORY ANALYSIS







OUTLINE

1. Introduction

2. Dataset

3. Proposed Approaches

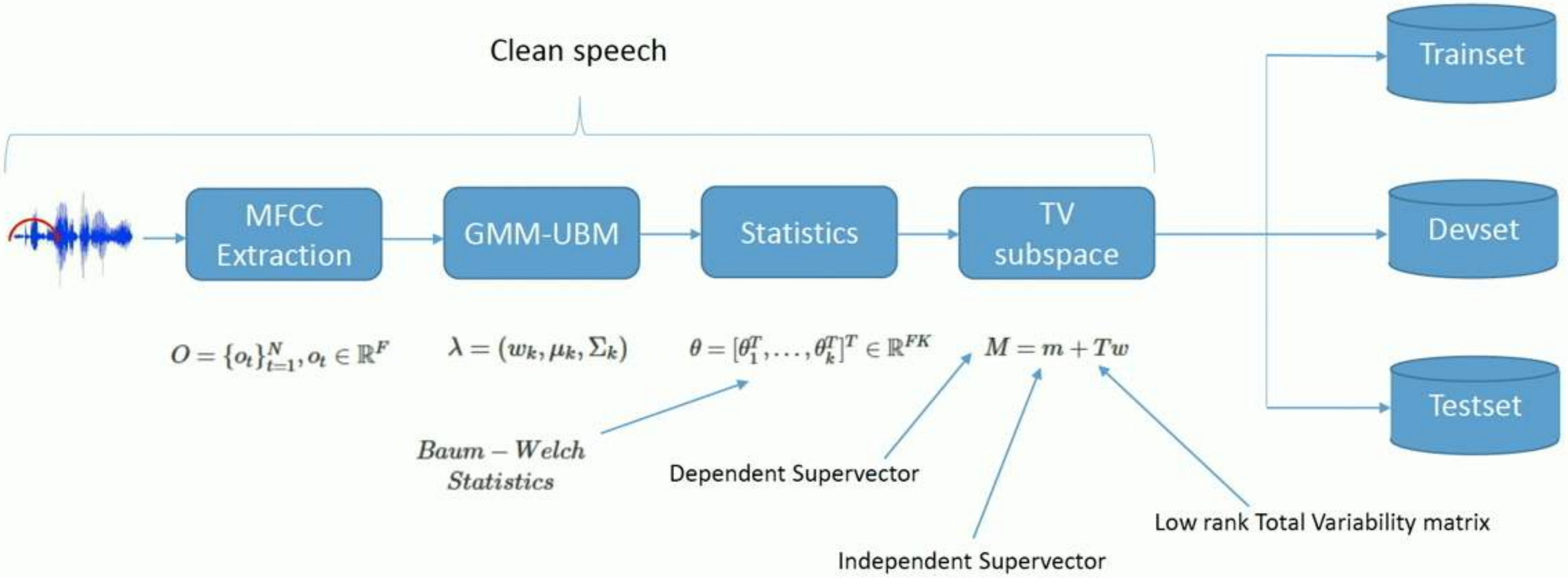
4. Results and Conclusion

I-VECTOR

- Widely used for recognition tasks involving speech:
  - Speaker Verification
  - Language Recognition
  - Emotion Recognition
  - Speech Quality?

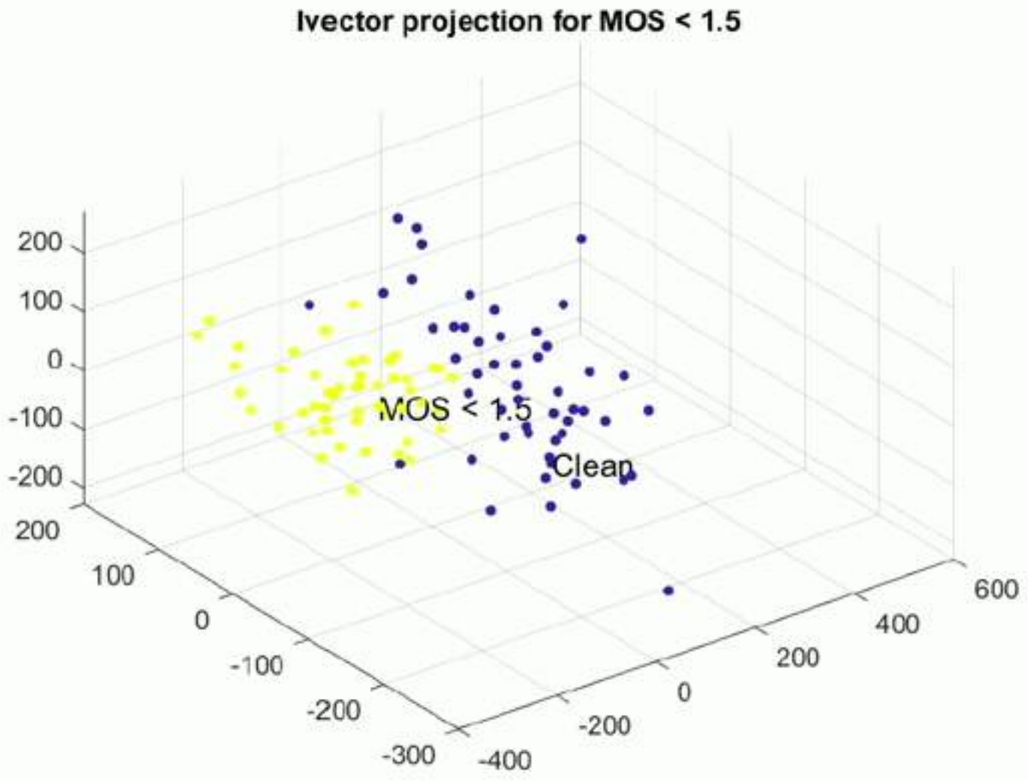
- I-vectors are known to carry speaker and channel variability

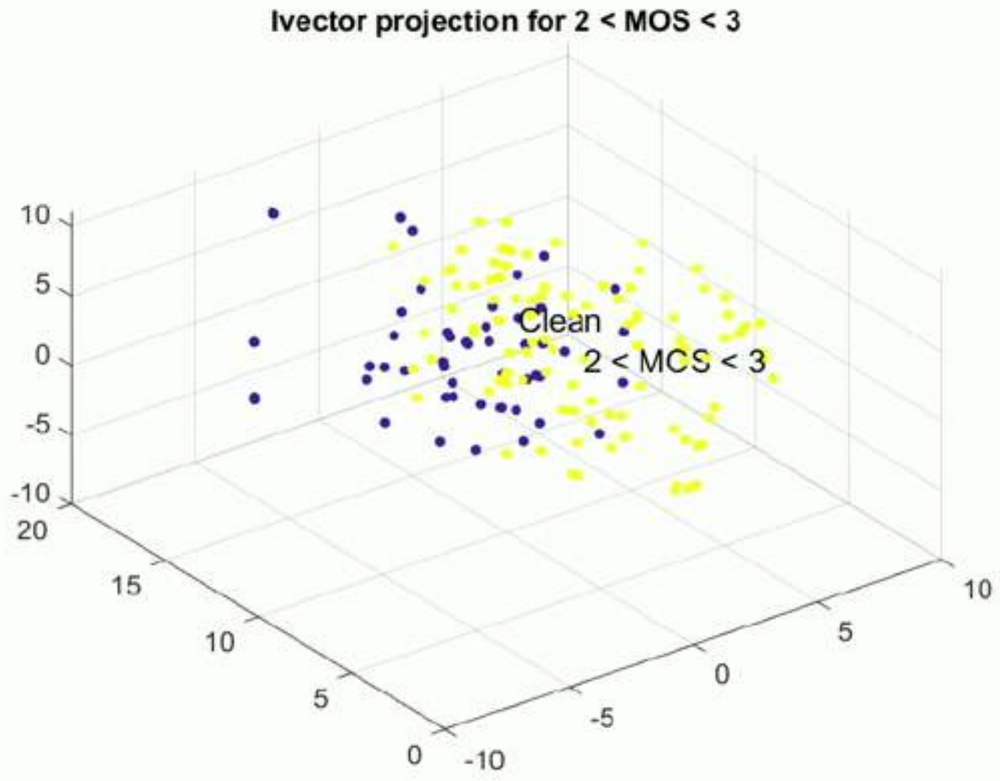
I-VECTOR



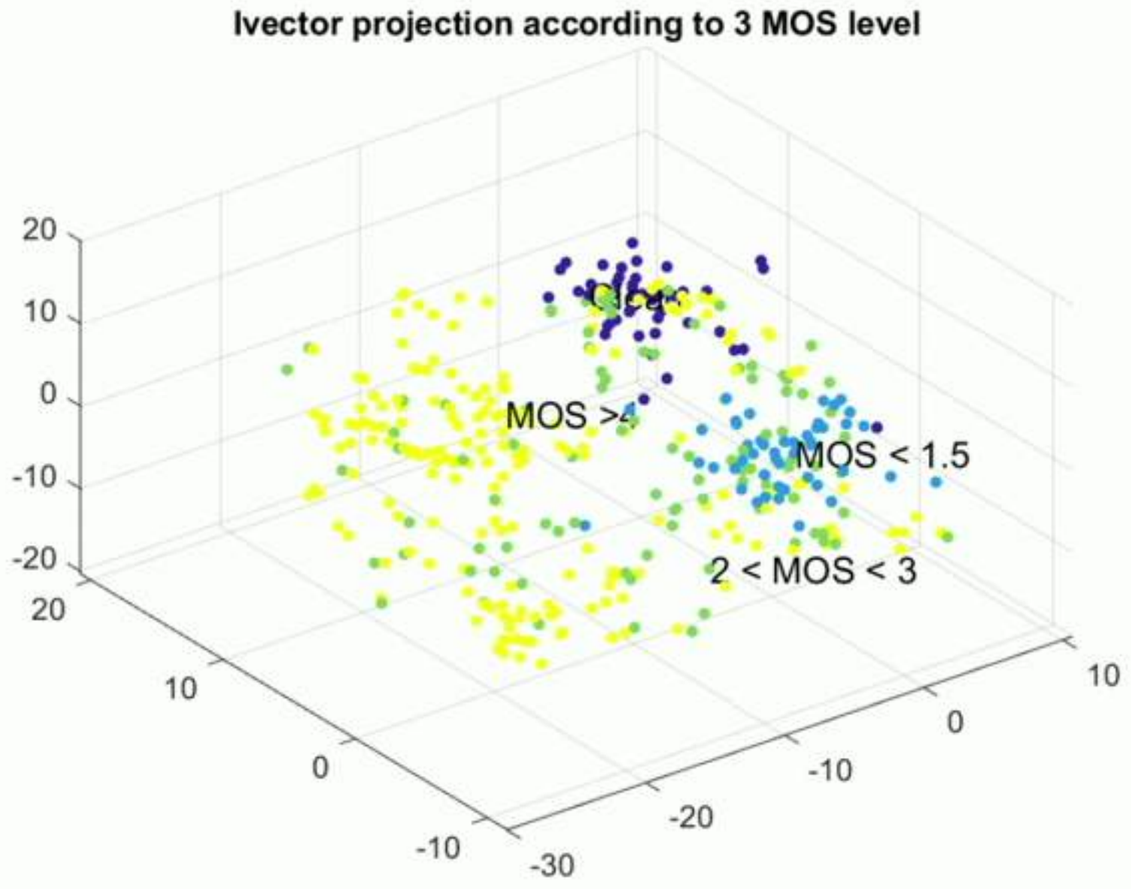
Dehak, Najim, et al. "Front-end factor analysis for speaker verification." *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011): 788-798.

I-VECTOR

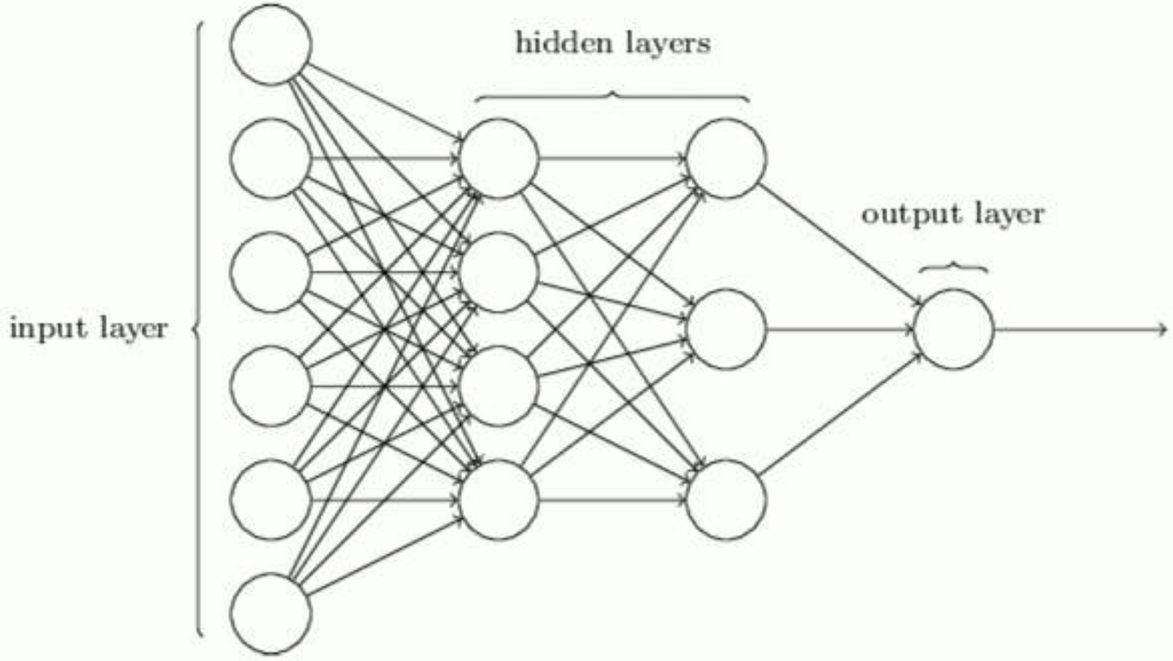

Ivector projection for MOS < 1.5
MOS < 1.5
Clean
200
100
0
-100
-200
-300
-400
600
400
200
0
-200


Ivector projection for 2 < MOS < 3
Clean
2 < MCS < 3
10
5
0
-5
-10
20
15
10
5
0

I-VECTOR


Ivector projection according to 3 MOS level
MOS >4
MOS < 1.5
2 < MOS < 3
20
10
0
-10
-20
20
10
0
-10
10
0
-10
-20
-30

I-VECTOR + DNN



400 input units

2 hidden layers
200 hidden units first layer
100 hidden units second layer
Relu as activation function
Dropout for regularization

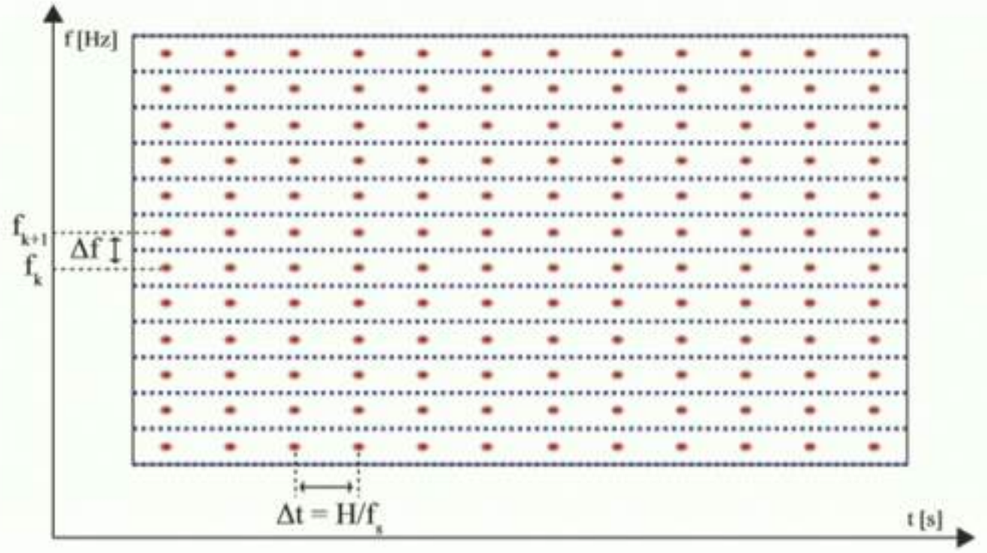
MOS prediction

CONSTANT Q CEPSTRAL COEFFICIENTS

- Recently used to detect natural and spoofed speech
- Also used in music signal processing
- **Psychoacoustic motivated**

CONSTANT Q CEPSTRAL COEFFICIENTS

$$Q = \frac{f_k}{\delta f}$$

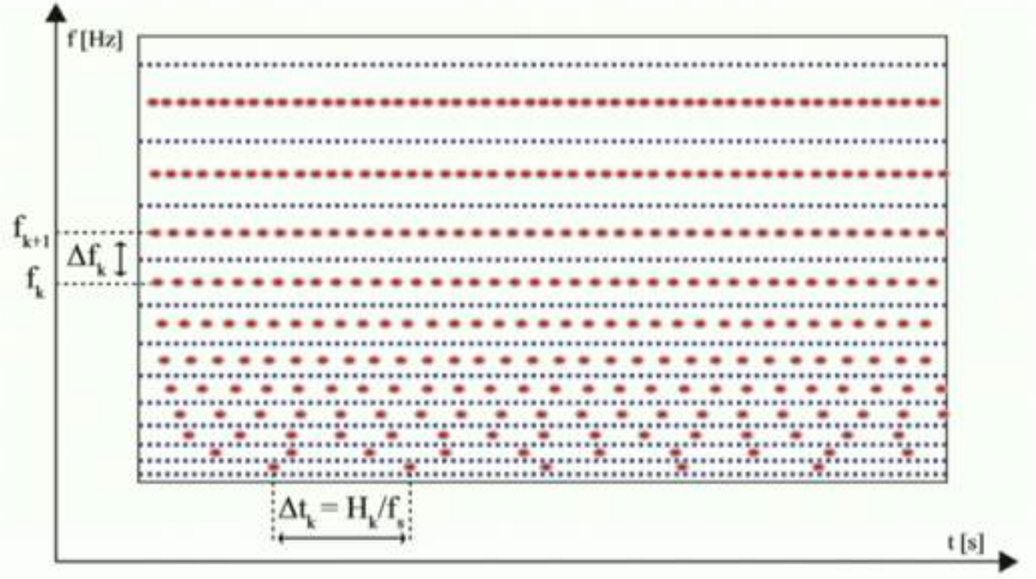


STFT

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.

CONSTANT Q CEPSTRAL COEFFICIENTS

$$Q = \frac{f_k}{\delta f}$$



CQT

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.

CONSTANT Q CEPSTRAL COEFFICIENTS

$$CQCC(p) = \Sigma log|X^{CQ}(l)|^2 cos\left[\frac{p(l-\frac{1}{2})\pi}{L}\right]$$

CONSTANT Q CEPSTRAL COEFFICIENTS

CNN (Q Cepstral Coefficients)

60x300 feature dimension
32 3x3 filters (2 layers)
2x2 Max Pooling
Dropout as regularizer
64 2x2 filters (2 layers)
2x2 Max Pooling
Dropout as regularizer
FC 64 hidden units
relu as activation function

CNN (Spectrum computed using constant Q transform)

240x300 feature dimension
32 3x3 filters (2 layers)
2x2 Max Pooling
Dropout as regularizer
64 3x3 filters (2 layers)
2x2 Max Pooling
Dropout as regularizer
FC 64 hidden units
relu as activation function

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.

OUTLINE

1. Introduction

2. Dataset

3. Proposed Approaches

4. Results and Conclusion

EVALUATION

- Evaluation parameters
  - Pearson's correlation
  - Mean Squared Error (MSE)
- Baselines
  - MOS as ground truth
  - SRMR, P563 and PESQ
- All results are from test set

EXPERIMENTAL RESULTS

	Pearson's correlation	MSE
I-vector (MLP)	0.79	0.22
Q Cepstral Coeficients (CNN)	0.79	0.23
Q Constant Spectrum (CNN)	0.85	0.16
PESQ	0.70	0.25
SRMR	0.60	0.31
P563	0.55	0.36
MEL features, FC (1024 x 4)	0.83	0.22
… + ELM (ELM stage)	0.83	0.17

CONCLUSION

- New audio quality dataset was created and labeled
- We explored the effect of mixed impairments on MOS
- Three approaches to estimate MOS were developed
- Our methods outperformed all of the baseline methods

NEXT STEPS

- Combine the I-vector framework with Q Cespstral features
- Test our algorithms on other datasets
- Explore more neural network architectures:
  - end-to-end DNN
  - DNN + ELM
  - CNN + ELM

THANK YOU!

- To Skype team for funding my internship: Johannes Gehrke and Scott Van Vliet
- To Chandan A. K. Reddy, Ross Cutler, Sriram Srinivasan for the help and fruitful discussions
- To Hannes Gamper and David Johnston for the help and consultations around neural networks
- To the Audio and Acoustcis Research Group in MSR Labs – Redmond for hosting my internship

Q & A

CONCLUSION

- New audio quality dataset was created and labeled
- We explored the effect of mixed impairments on MOS
- Three approaches to estimate MOS were developed
- Our methods outperformed all of the baseline methods

DATA EXPLORATORY ANALYSIS