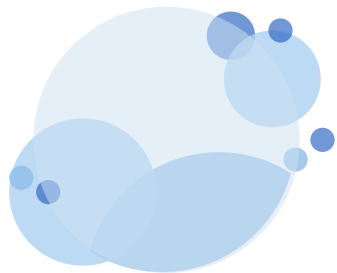




DIVER:

machine learning-based Audio quality Assessment
without additional far-end reference signal

Hsiao-Tzu (Anna) Hung 洪筱慈
Internship Presentation



Outline

1. Introduction
2. NISQA - Test Result
3. NISQA - Methods
4. Root Cause Analysis
5. User Interface
6. Future Work

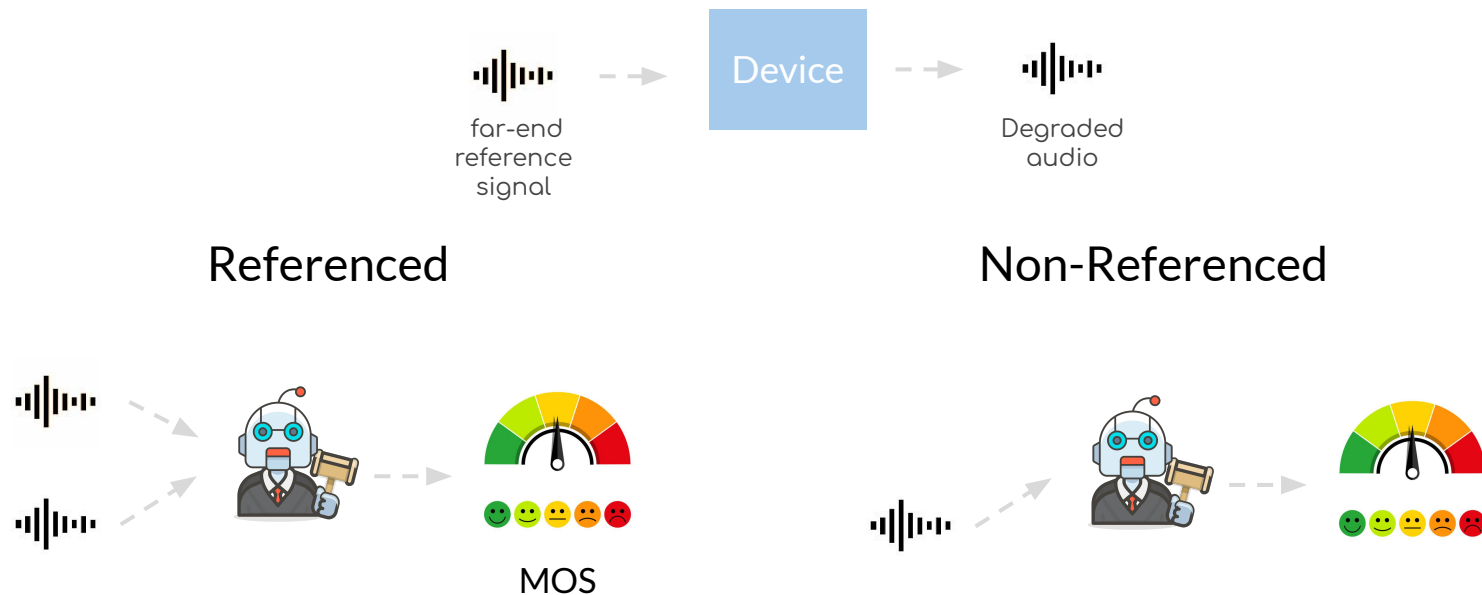
The user interface can be downloaded from here: For [Windows](#) or for [Mac](#).

It currently support the MOS estimator and version (1) of distortion recognition model(page 17.)

All the wave files in this slides can be found in [this folder](#).

1. Introduction

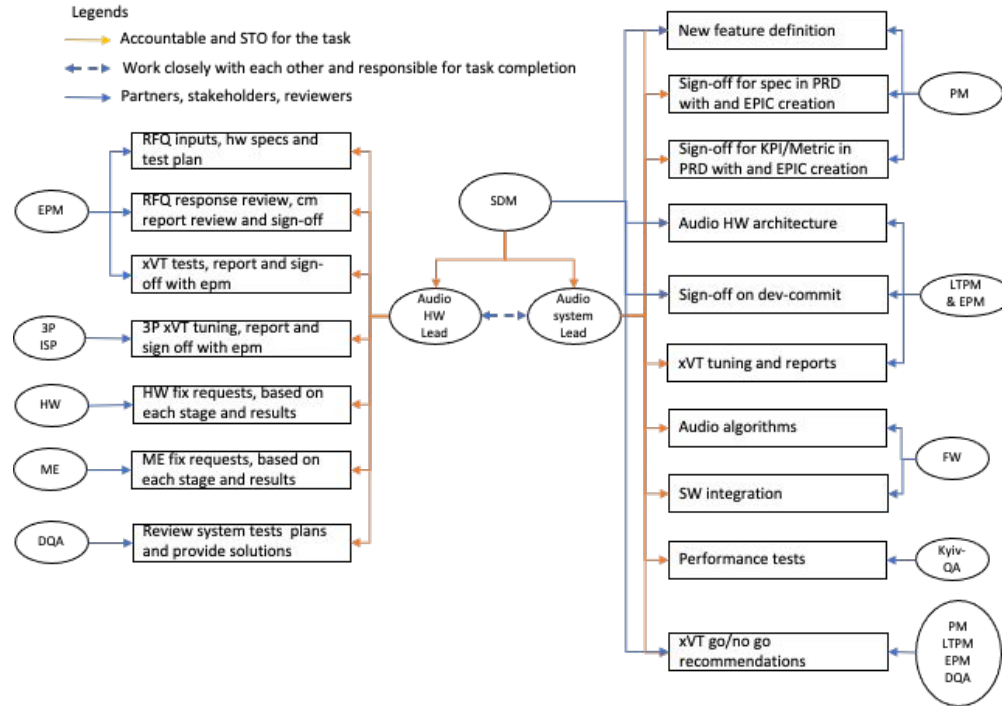
Audio quality assessment



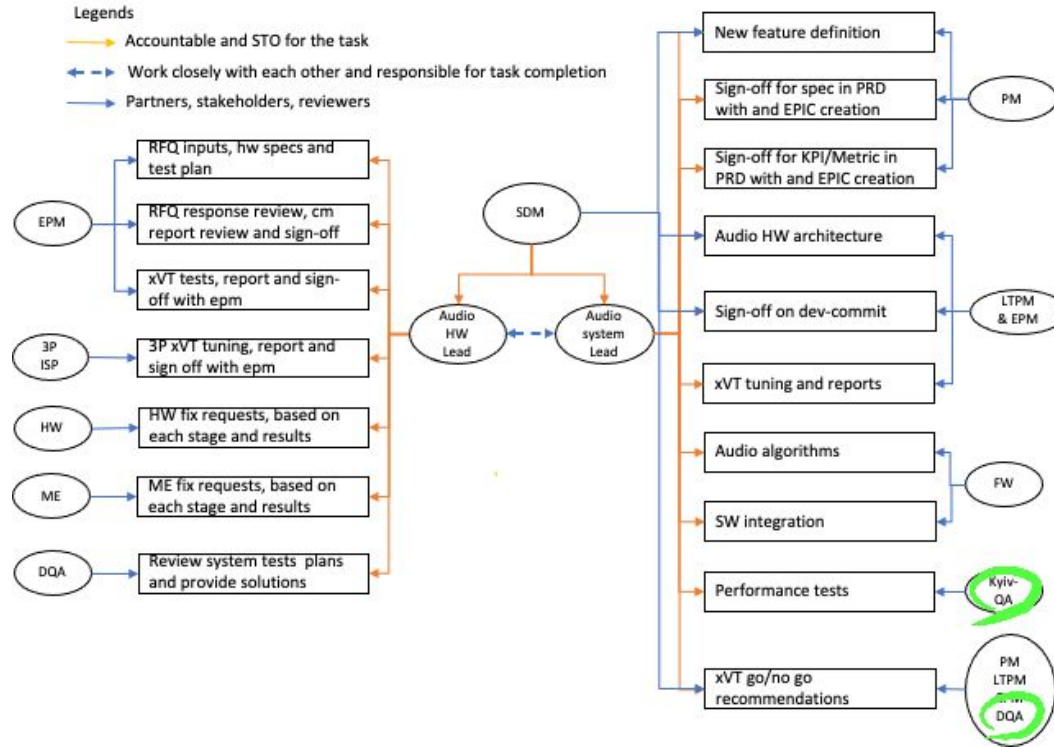
Sound quality could be efficiently analyzed via **limited information** and **cheap** equipment.

Referenced way ACQUA, POLQA, PESQ		Non-referenced way NISQA, MOSNet, MBNet
Far-end referenced audio needed. Unavailable for some testing scenarios	Flexibility	Far-end referenced audio is no needed
Need to be tested by machines in the lab and operated by engineers	Automation	Calculated by Python scripts, and can be integrated into automatic testing pipeline
ACQUA : USD 0.3 M POLQA : USD 4,900*	Cost	No cost for internal use
Need to go to the lab.	Time	Several seconds on personal laptop

Where can we use it?

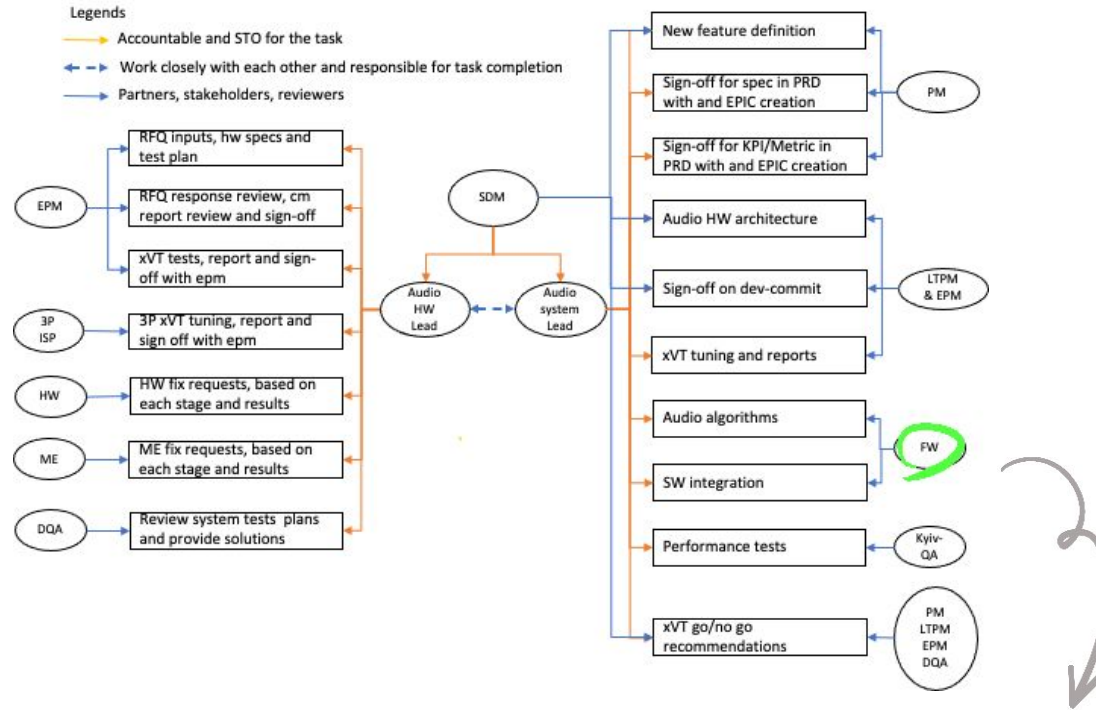


Where can we use it?



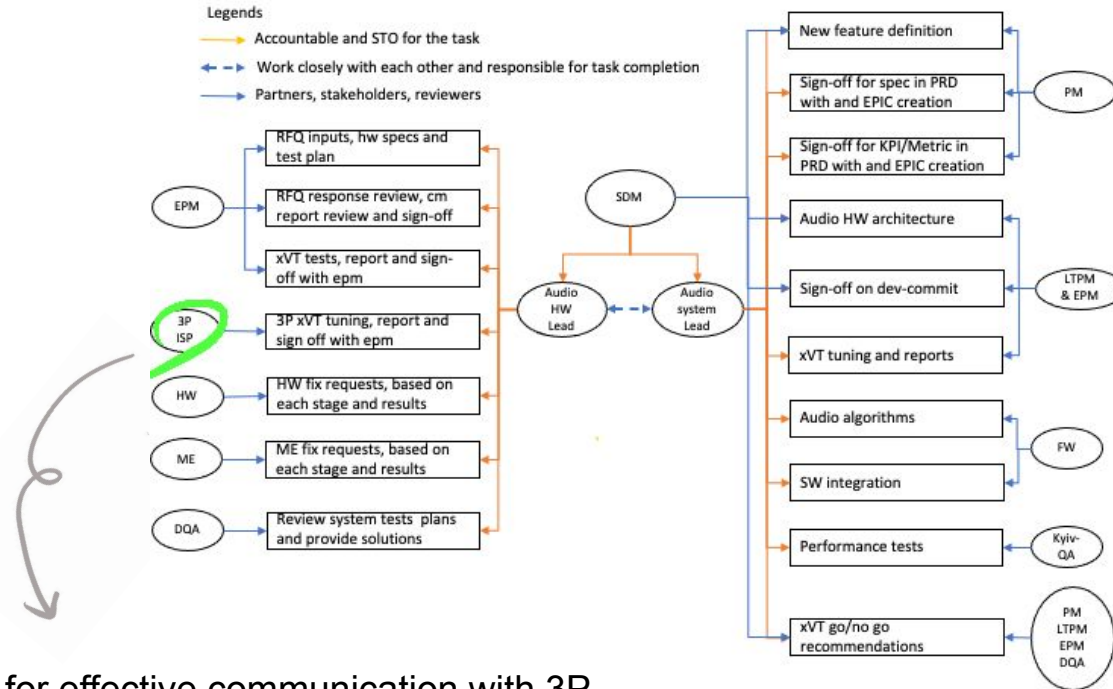
An efficient and objective way to observe quality during early stage of development

Where can we use it?



Help firmware engineer to detect regading while turning the system

Where can we use it?






A metric for effective communication with 3P

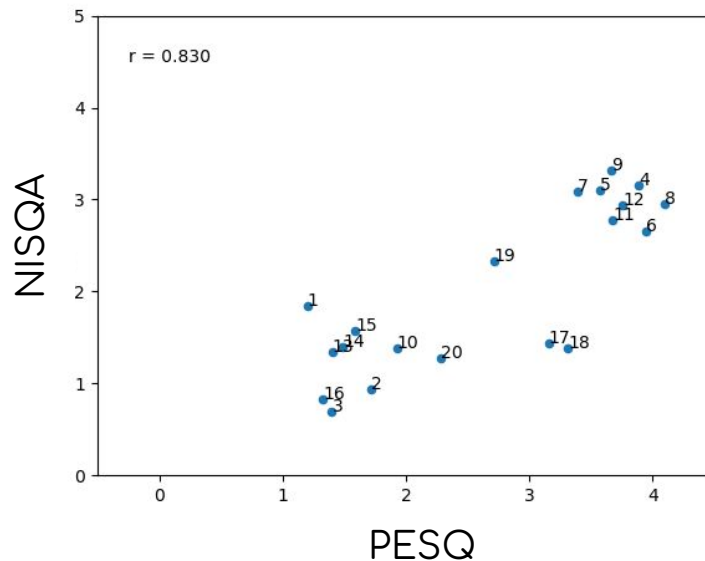
2. Result

A. Compare to referenced approach: PESQ*

Device*: FLC, ironman, hazelnut and Lunar

The Pearson correlation coefficient is .83 with a p-value of .042, which is a strong positive correlation.

		NISQA	PESQ
	No.3	0.696	1.400
	No.10	1.378	1.928
	No.4	3.150	3.887



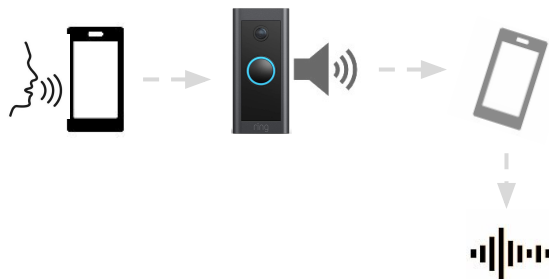
*Refer to Appendix B.1 to see the difference between PESQ and POLQA.

*Refer to the [conference page](#) to listen to the audios.

* The range of PESQ is [-0.5, 4.5], and the range of NISQA is [0, 5]

B. Test on latest products

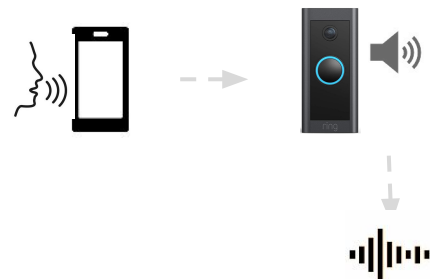
A.speaker playback



B.Microphone recording












C.2-way audio, single-talk



A.speaker playback

B.MIC recording

C. 2-way audio, single-talk

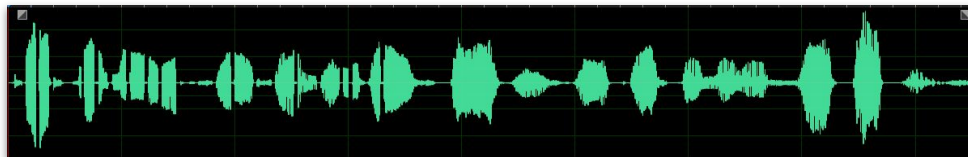
	A.speaker playback	B.MIC recording	C. 2-way audio, single-talk
GC	 0.91700	 1.88763	 3.25706
S'more	 1.00677	 1.86236	 4.01799
Jellyfish	 1.35994	 2.09023	 1.49201

Encountered packet loss :



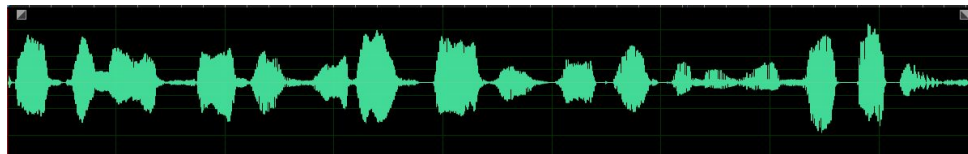
1.49201

Partial packet loss :



1.66996

No packet loss :



4.05084

	A.speaker playback	B.MIC recording	C. 2-way audio, single-talk
GC	0.91700	2.08465	3.25706
S'more	1.00677	2.63851	4.01799
Jellyfish	1.35994	1.85219	4.05084

3. Method

NISQA is a piece of work that got accepted by Interspeech 2021:

NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets

Gabriel Mittag¹, Babak Naderi¹, Assmaa Chehadi¹, Sebastian Möller^{1,2}

¹Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany

²Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany

`first.last@tu-berlin.de`

According to the Amazon science website*:

Interspeech is a technical conference focused on speech processing and application, emphasizing interdisciplinary approaches addressing all aspects of speech science and technology, ranging from basic theories to advanced applications.

Alexa speech, Alexa TTS, and Amazon Chime teams also have 30+ papers got accepted this year.

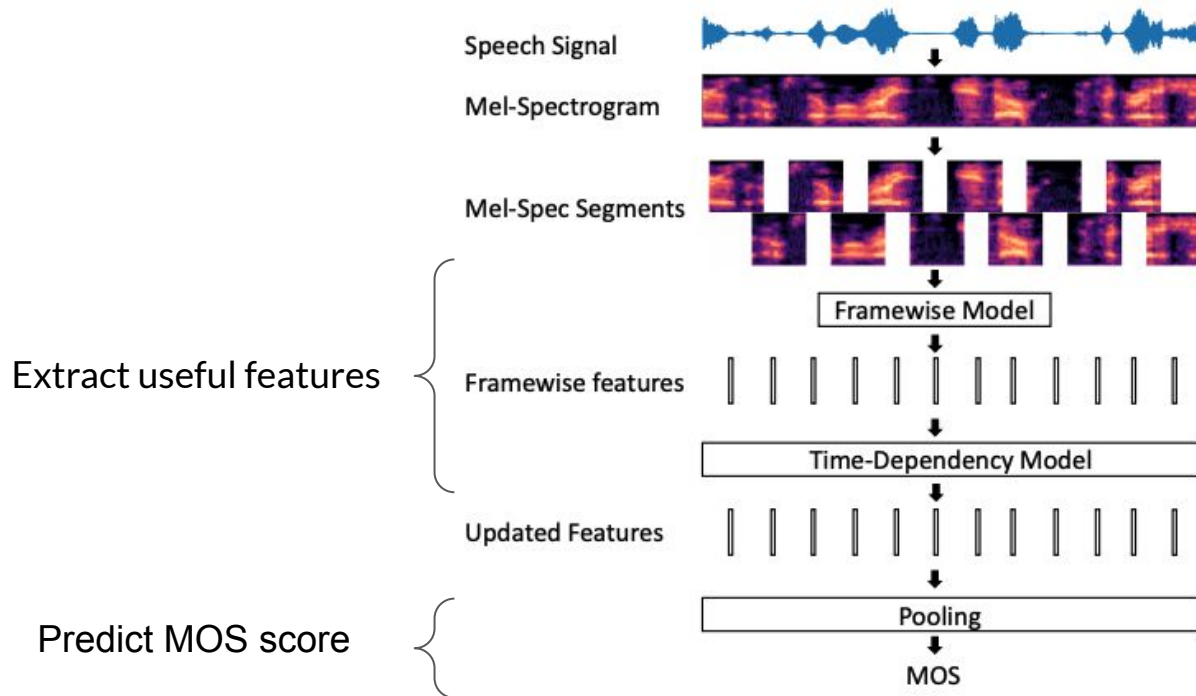
*<https://www.amazon.science/conferences-and-events/interspeech-2021>
<https://arxiv.org/pdf/2104.09494.pdf>

Training data

- The size of the training data is **72,903** VoIP audio files. (simulated or from skype, wechat)
- Language: German, Australian English, British English
- There are both men and women speakers in the speech data
- Distortion types:
 - Additive white Gaussian noise
 - Signal correlated MNRU noise.
 - Randomly sampled noise clips taken from the DNS-Challenge dataset
 - Lowpass / highpass / bandpass / arbitrary filter with random cutoff frequencies
 - Amplitude clipping
 - Speech level changes
 - Codecs in all available bitrate modes: AMR-NB, AMR-WB, G.711, G.722, EVS, Opus
 - Codec tandem and triple tandem
 - Packet-loss conditions with random and bursty patterns.
 - Combinations of the different distortions

For more information, please refer to the [introduction page](#) and original [dataset Wiki](#)

Model Structure



Results reported in the original NISQA paper:

- For the real live audio data, NISQA outperformed POLQA.
- The correlation between NISQA and crowdsourcing label is 0.82 and 0.90
- The RMSE is about 0.35 and 0.40, so we might bear in mind that the output is a bit random. (can test on our device data, use our data to calibrate the model)

Table 4: *Per-condition validation and test results of the overall quality in terms of PCC and RMSE after first-order mapping.*

Dataset	Scale	Lang	Con	Files	NISQA		P563		ANIQUE+		WAWEnets		POLQA		DIAL		VISQOL	
					r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE
103.ERICSSON	SWB	se	54	648	0.85	0.38	0.36	0.66	0.54	0.60	0.28	0.68	0.87	0.34	0.78	0.45	0.26	0.69
104.ERICSSON	NB	se	55	660	0.77	0.47	0.64	0.57	0.68	0.55	0.13	0.74	0.91	0.31	0.76	0.49	0.39	0.69
203.FT.DT	SWB	fr	54	216	0.92	0.36	0.68	0.69	0.47	0.82	0.64	0.72	0.91	0.38	0.79	0.57	0.59	0.75
303.OPTICOM	SWB	en	54	216	0.92	0.33	0.85	0.44	0.71	0.59	0.43	0.76	0.93	0.31	0.71	0.59	0.42	0.76
403.PSYTECHNICS	SWB	en	48	1152	0.91	0.36	0.81	0.50	0.77	0.54	0.78	0.53	0.96	0.24	0.92	0.34	0.73	0.57
404.PSYTECHNICS	NB	en	48	1151	0.77	0.39	0.82	0.35	0.74	0.41	0.14	0.61	0.86	0.31	0.67	0.46	0.55	0.51
503.SWISSQUAL	SWB	de	54	216	0.92	0.34	0.71	0.62	0.61	0.70	0.59	0.71	0.94	0.29	0.85	0.46	0.65	0.67
504.SWISSQUAL	NB	de	49	196	0.92	0.37	0.83	0.50	0.79	0.56	0.54	0.77	0.87	0.45	0.73	0.63	0.60	0.73
603.TNO	SWB	nl	48	192	0.89	0.44	0.83	0.53	0.69	0.69	0.59	0.77	0.95	0.29	0.86	0.48	0.47	0.84
ERIC.FIELD.GSM.US	NB	en	372	372	0.79	0.36	0.42	0.54	0.17	0.58	0.60	0.47	0.75	0.39	0.71	0.42	0.51	0.51
HUAWEL2	NB	zh	24	576	0.98	0.21	0.93	0.35	0.79	0.59	0.63	0.75	0.94	0.32	0.89	0.44	0.97	0.24
ITU.SUPPL23.EXP1o	NB	en	44	176	0.92	0.31	0.90	0.34	0.98	0.15	0.73	0.53	0.91	0.32	0.91	0.33	0.86	0.39
ITU.SUPPL23.EXP3d	NB	ja	50	200	0.92	0.27	0.93	0.26	0.97	0.17	0.68	0.50	0.85	0.36	0.84	0.36	0.79	0.41
ITU.SUPPL23.EXP3o	NB	en	50	200	0.91	0.30	0.91	0.30	0.98	0.15	0.79	0.45	0.88	0.35	0.87	0.36	0.78	0.45
TUB.AUS	FB	en	50	600	0.91	0.21	0.62	0.40	0.65	0.39	0.70	0.36	0.88	0.24	0.73	0.35	0.63	0.40
TUB.LIKE	SWB	de	8	96	0.98	0.25	0.85	0.60	0.85	0.61	0.59	0.93	0.99	0.16	0.89	0.53	0.81	0.67
NISQA.VAL.LIVE	FB	en	200	200	0.82	0.40	0.42	0.64	0.51	0.61	0.36	0.66	0.67	0.52	-0.22	0.69	0.66	0.53
NISQA.VAL.SIM	FB	en	2500	2500	0.90	0.48	0.45	0.99	0.54	0.93	0.30	1.05	0.86	0.56	0.36	1.03	0.78	0.69
NISQA.TEST.P501	FB	en	60	240	0.95	0.31	0.72	0.67	0.73	0.66	0.80	0.59	0.95	0.30	0.80	0.59	0.80	0.58
NISQA.TEST.NSC	FB	de	60	240	0.97	0.23	0.69	0.67	0.62	0.74	0.78	0.59	0.93	0.35	0.79	0.57	0.78	0.59
NISQA.TEST.FOR	FB	en	60	240	0.95	0.26	0.52	0.71	0.54	0.70	0.81	0.49	0.92	0.33	0.75	0.55	0.68	0.61
NISQA.TEST.LIVETALK	FB	de	58	232	0.90	0.35	0.70	0.58	0.56	0.68	0.66	0.61	N/A	N/A	N/A	N/A	N/A	N/A

5. Root cause analysis

Data

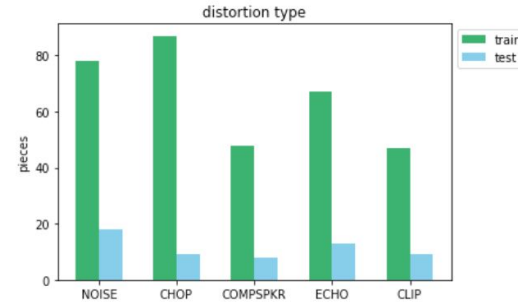
Because of the lack of GPU, first train on a small dataset and by simple algorithm.

Dataset: TCD-VoIP[1] dataset

- Train: 327 samples.
- Test: 57 samples

TABLE II: Degradations and Parameters used in TCD-VoIP

Degradation	Conditions	Parameters	Range
Chop	20	Rate	0-6 chops/s
		Period	0.02-0.04 s
Clip	10	Mode	Insert, Delete, Overwrite
		Multiplier	1-55
Competing Speaker	10	Gender code	1-5
		SNR	10-50 dB
Echo	20	Alpha	0-0.5
		Delay	0-220 ms
Noise	20	Noise Type	Car, Street, Office, Babble
MNRUs	4	SNR	5-55 dB
		SNR (Q)	48, 36, 24, 12



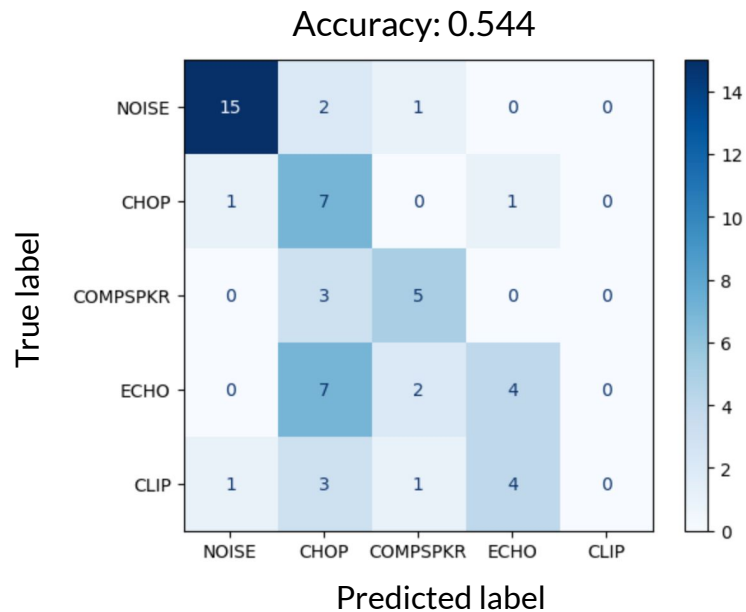
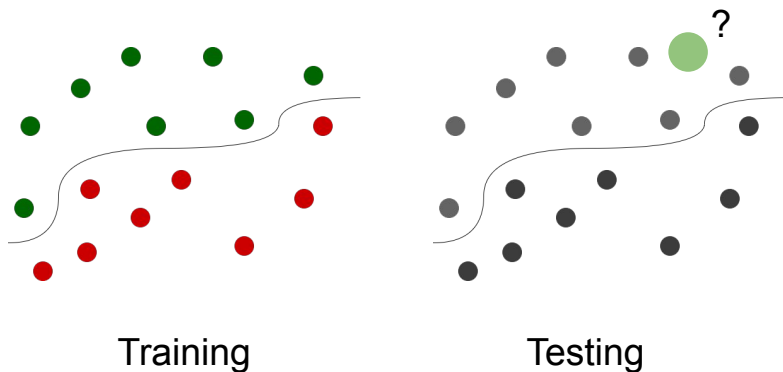
[1] TCD-VoIP, a Research Database of Degraded Speech for Assessing Quality in VoIP Applications

<https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1156&context=scschcomcon>

5. Root cause analysis

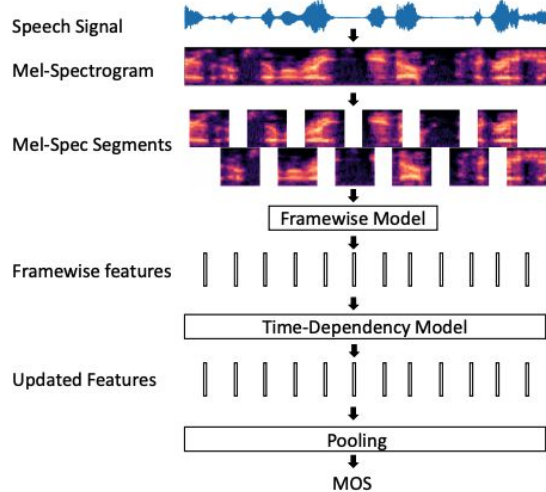
Method & Result(1): ML-based method

Waveform \rightarrow STFT \rightarrow Logistic regression

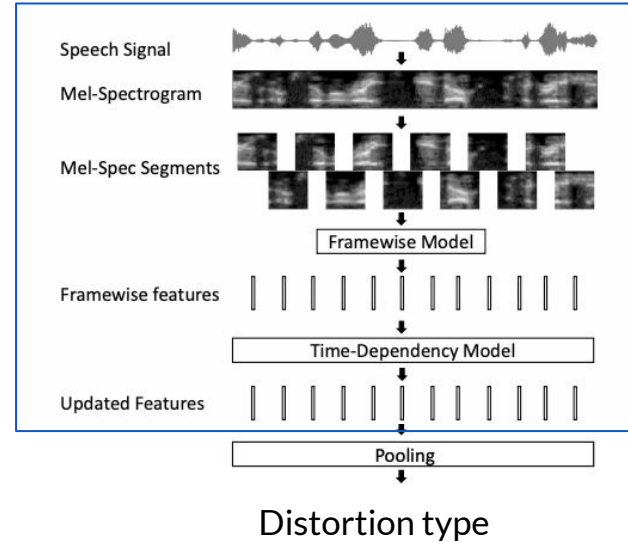


5. Root cause analysis

Method & Result(2): NISQA-based method



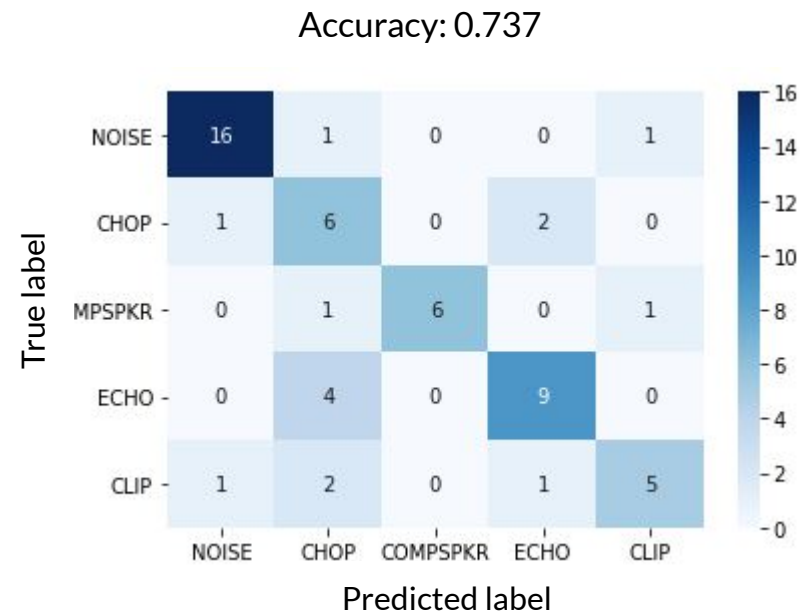
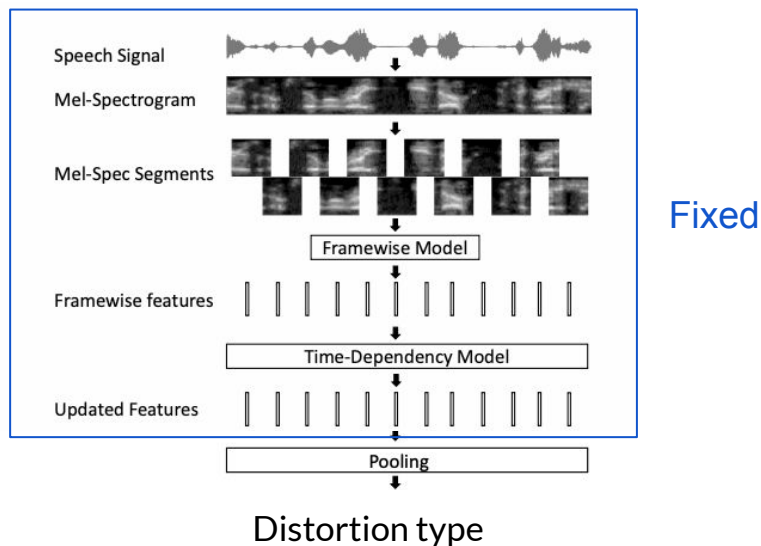
Original NISQA architecture for estimating MOS



Modified NISQA architecture for predicting distortion types

5. Root cause analysis

Method & Result(2): NISQA-based method

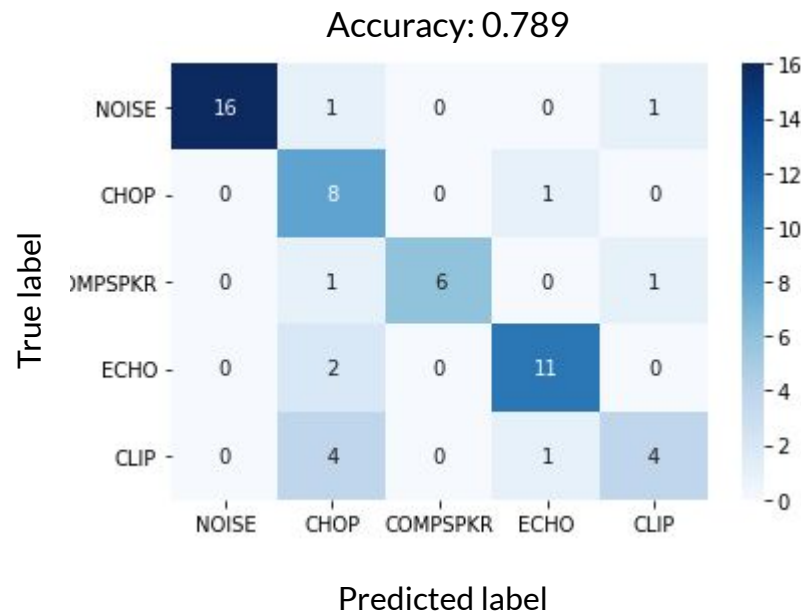


Modified NISQA architecture for predicting distortion types

5. Root cause analysis

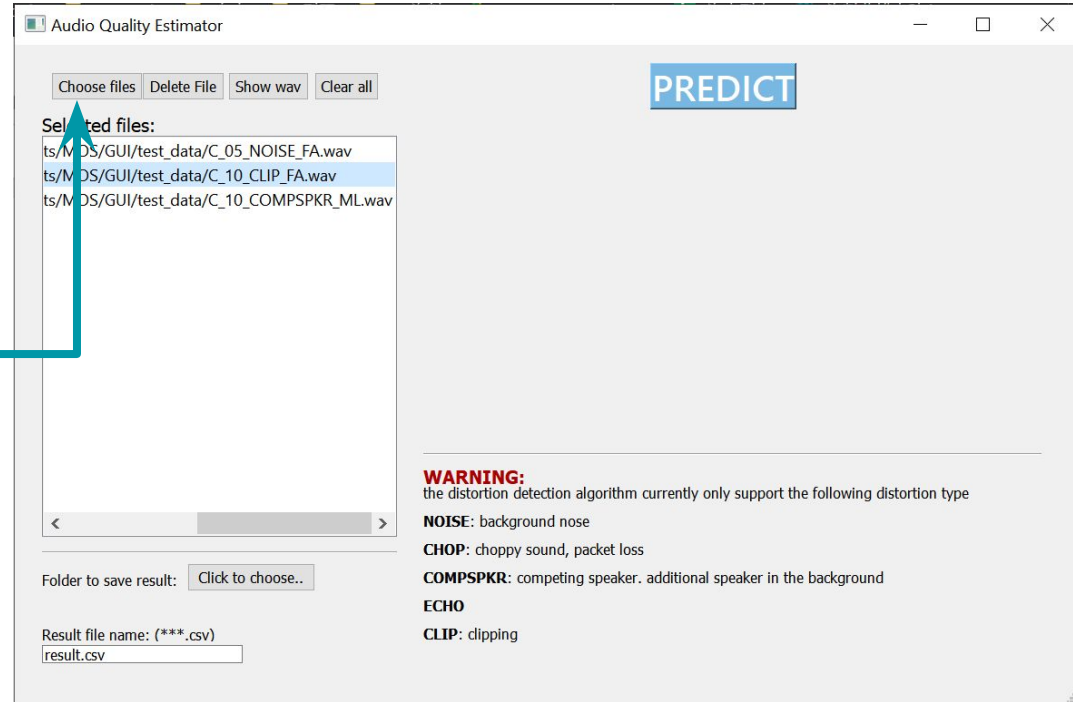
Method & Result(3): NISQA-based method + more data

- Utilize data from NISQA corpus (real data from VoIP communication)
 - Packet loss: + 80 samples
 - Clipping: + 24 samples



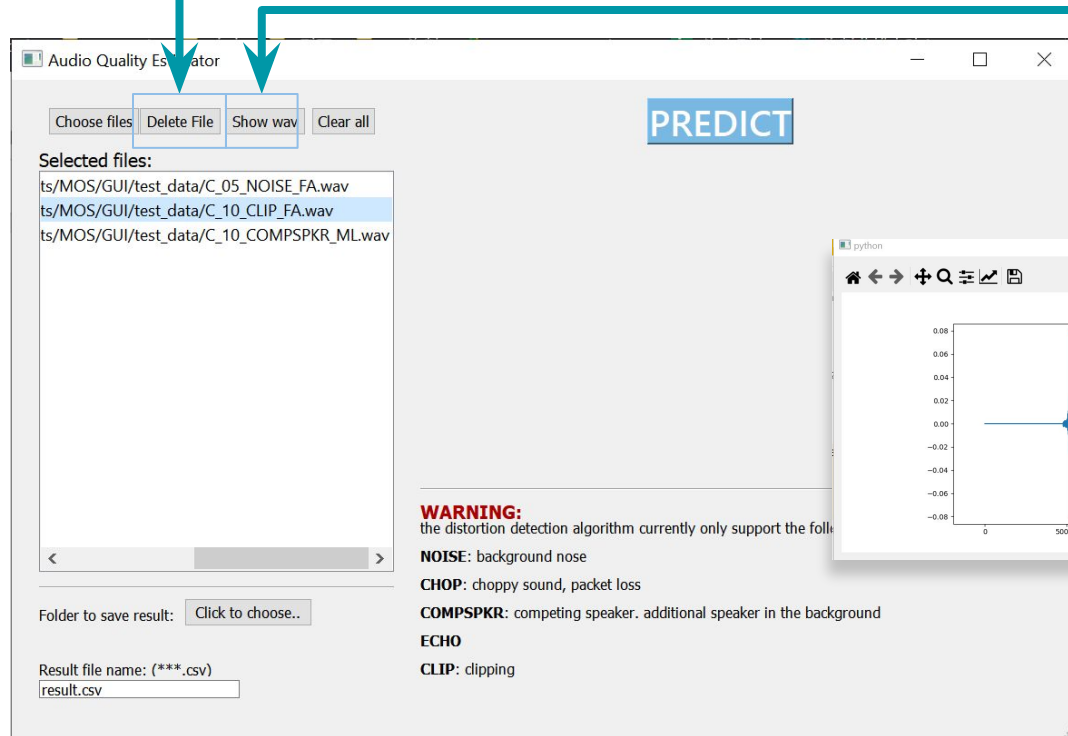
6. User Interface

Choose the file/files you want to analyze

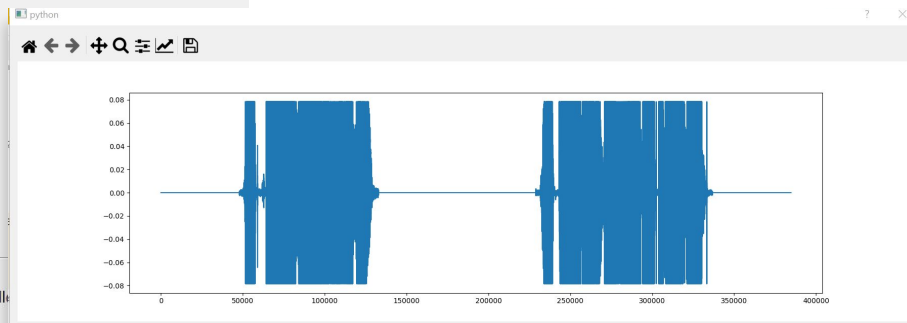


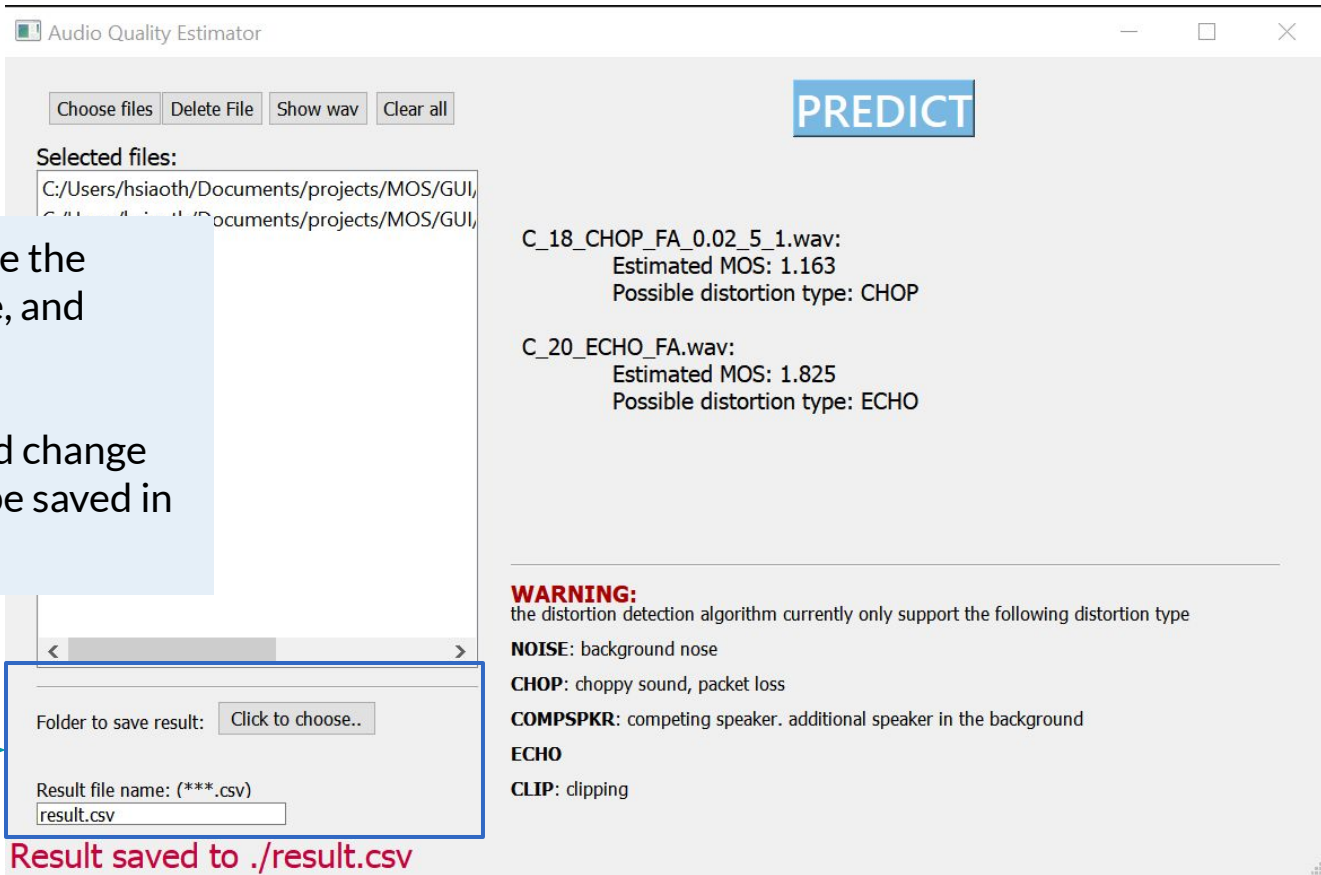
The user interface can be downloaded from here: For [Windows](#) or for [Mac](#).
It currently support the MOS estimator and version (1) of distortion recognition model(page 17.)

You can select an unwanted file and delete it



Or select a file to display the waveform





Choose a folder to save the estimate result csv file, and decide the name of it.

If you don't choose and change anything, the file will be saved in this program folder.

Audio Quality Estimator

Choose files Delete File Show wav Clear all

Selected files:
C:/Users/hciantn/Documents/projects/MOS/GUI/objects/MOS/GUI/

PREDICT

C_18_CHOP_FA_0.02_5_1.wav:
Estimated MOS: 1.163
Possible distortion type: CHOP

C_20_ECHO_FA.wav:
Estimated MOS: 1.825
Possible distortion type: ECHO

WARNING:
the distortion detection algorithm currently only support the following distortion type

NOISE: background noise

CHOP: choppy sound, packet loss

COMPSPKR: competing speaker. additional speaker in the background

ECHO

CLIP: clipping

Folder to save result: Click to choose..

Result file name: (*.csv)
result.csv

Result saved to ./result.csv

Click the “**PREDICT**” button, it will start estimating the files you select, and show the results in the block below.

For each file, there will be 2 results:

- Estimated MOS: (1-5)
- Possible distortion type (among the supported types listed below.)

Conclusion

- **Non-referenced audio quality assessment** is more flexible, efficient and cheaper than referenced method.
- Found and built a non-referenced method called **NISQA**. It is **highly correlated** with PESQ using recording of our devices.
- NISQA is sensitive to degradation, such as packet-loss, choppy speech, clipping.
- Current root cause analysis tool is still not stable enough and need more data for training.

Future work

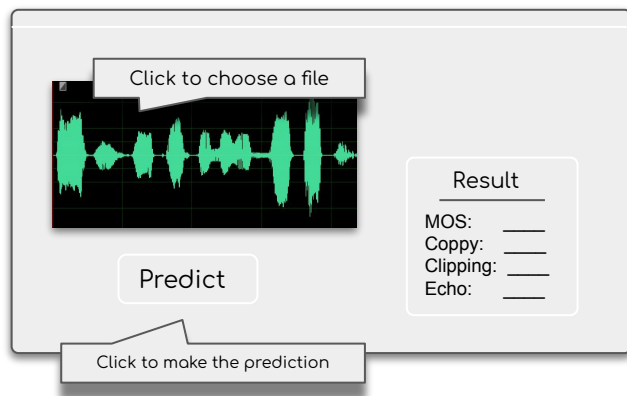
Root cause analysis:

- a. Simulate more distortion data to improve the accuracy.
- b. Build a few-shot learning system so that we can increase the type of distortion easily in the future.
 - i. Simulate more data (1000+ per type) of the existing types, for example: clipping, echo.
 - ii. Use the data to build a stronger recognition model.
 - iii. Collect some data for a new wanted distortion type. (10 samples is enough).
 - iv. Use few-shot learning training process to train on the new distortion type data.

(This is a method that has achieved good performance in image recognition.

Future work

1. User friendly interface
2. Root cause analysis: general degradation detection
3. Improve **Robustness**: (8/27)
 - a. Solve the time dependency problem: data augmentation, random shuffle in the final layer
 - b. More stable prediction -> make the deviation of the same situation smaller
4. **Mapping** NISQA and ACQUA
5. Build up AWS EC2 GPU environment



1.66996 \approx 2

2.3836 \approx 2

One minute about Machine Learning...

How did the model “learn”?

- Objective-driven(考試領導教學)
- Data-driven(題庫做好做滿)

(actually ACQUA is also a learning-based method, but they didn't share their detailed method)

- **Audio data:**

- Simulated speech distortions:

- Packet-loss
 - Bandpass filter
 - Different codecs
 - Clipping
 - Background noises

- Live (e.g. mobile phone, Zoom, Skype, WhatsApp) conditions

- **MOS label:** crowdsourcing annotation

What's the strength of the model?

The relationship between data and answer are complex, and cannot use simple function to approach it.

Machine Learning
 \approx Looking for Function

- Speech Recognition

$f(\text{audio waveform}) = \text{"How are you"}$

- Image Recognition

$f(\text{cat image}) = \text{"Cat"}$

- Playing Go

$f(\text{Go board state}) = \text{"5-5" (next move)}$

What can't the model directly do?

- The **type** of the testing data is too different from the training data.
- The testing **task** is different from the training phase.

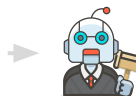
Training



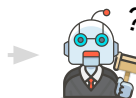
Testing



Training



Testing



?

Does clipping
Occur in the speech?

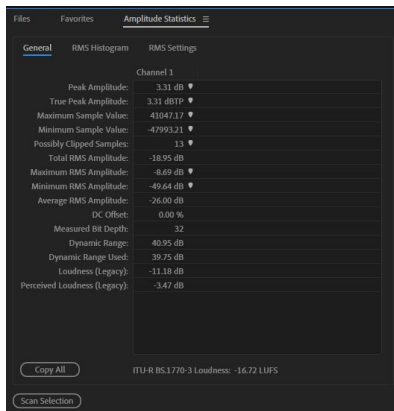
Appendix A: Other works

Automatic calibration

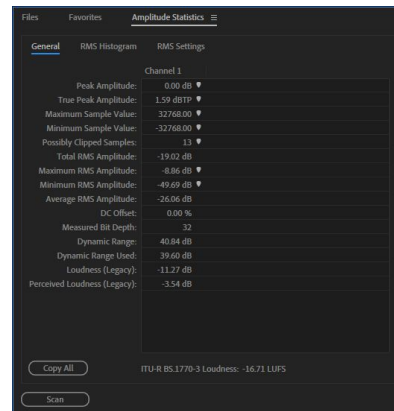
It's common that some calibration processes are needed before taking the quality test. Originally, we will use the Audition software to do the Active Speech Level calibration by changing the average RMS Amplitude to -26dB.

I use an open source python too called pydub to do this automatically, so that we don't need to calibrate the file manually before making MOS prediction.

Since Audition don't share how they do the amplification, we can only compare the statistics of the calibrated results. The following figures show the statistics of the calibrated speech file using (A) Audition (B) pydub. We can see that the calibrated result of pydub is slightly different from Audition, but the MOS scores of the two calibrated speeches are almost the same, and also they sound the same to me. So I think we are able to use it.



(A) Audition: 2.70975

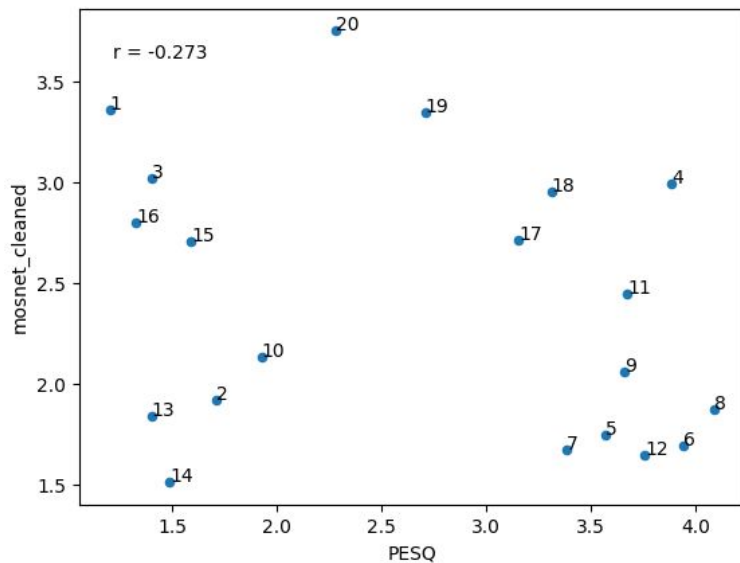


(B) pydub, MOS=2.70342

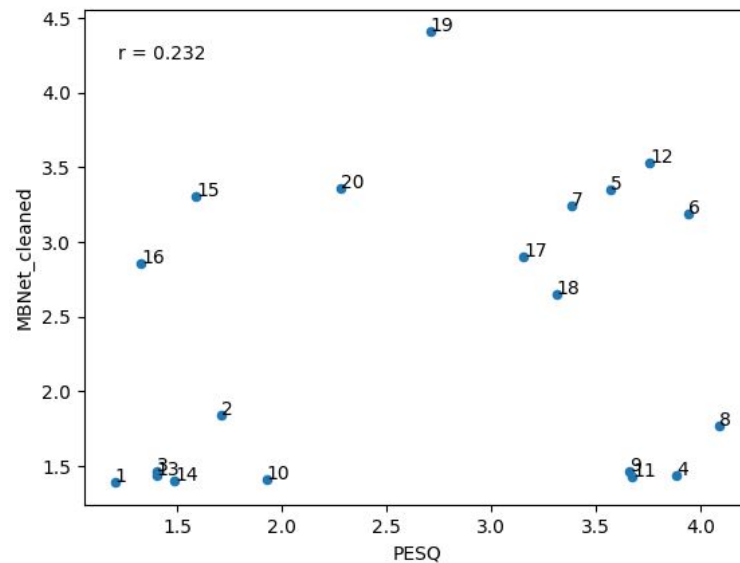
Other referenced methods (correlation are too low to use it directly)

These methods trained the model on speeches generated by voice conversion models, so they are not suitable for us.

MOSNet[1]



MBNet[2]

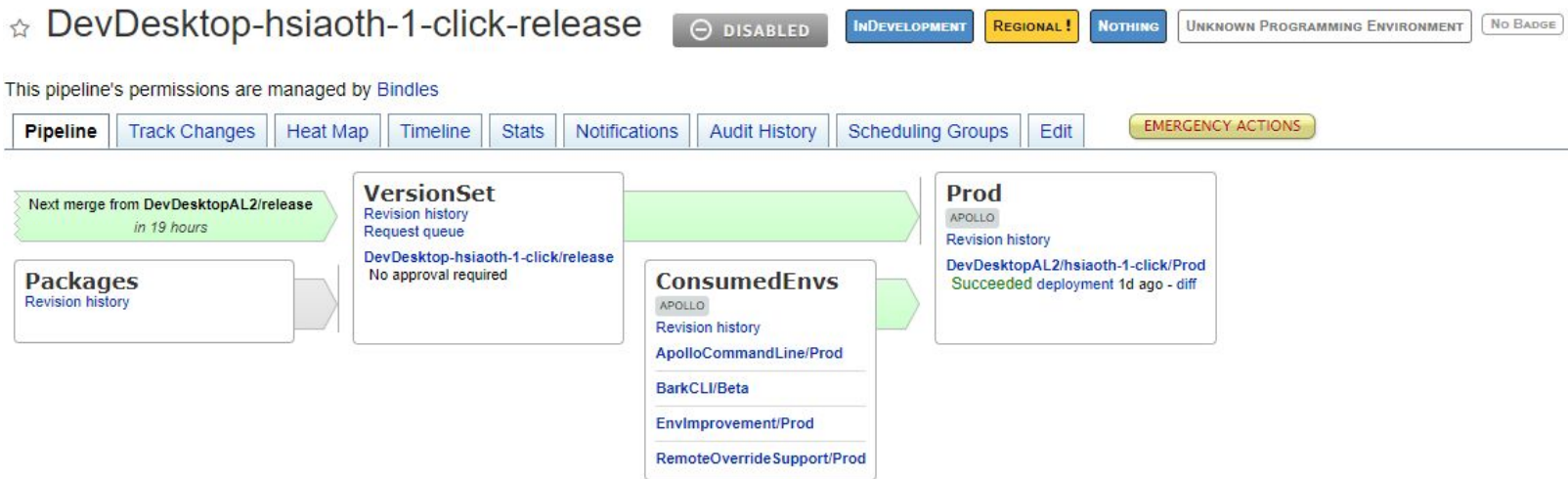


Survey on the state-of-the-art speech quality assessment works and speech datasets

- Speech-related dataset:
 - <https://betterprogramming.pub/assessing-audio-quality-with-deep-learning-f66d1761f938>
 - <https://towardsdatascience.com/a-data-lakes-worth-of-audio-datasets-b45b88cd4ad>
 - ITU-T Rec. P.Sup23 (The coded speech database is delivered on three CD-ROMs), [DOC](#)
 - TUT-T p.862: https://github.com/denniguse/ITU-T_pesq
 - NISQA: <https://github.com/gabrielmittag/NISQA/wiki/NISQA-Corpus>
- Non-referenced audio quality assessment
 - [DNN No-Reference PSTN Speech Quality Prediction](#), TU Berlin, Microsoft Corp. , open source dataset: TBD
 - [Bias-Aware Loss for Training Image and Speech Quality Prediction Models from Multiple Datasets](#)
 - [Development of a Speech Quality Database Under Uncontrolled Conditions](#)
 - [NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets](#)
 - ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric: <https://arxiv.org/pdf/2004.09584.pdf>

Build cloud desktop for using AWS EC2

Already successfully setup the cloud desktop. The next step will be to deal with the settings about using AWS EC2 GPU.



Appendix B: Reference

B.1 PESQ v.s. POLQA

PESQ versus POLQA Overview

	PESQ	POLQA
Acoustic measurements	☹ Not easy	☺
Correct scoring with high background noise	☹	☺
AMR vs EVRC codec comparison	☹	☺
Representative scoring of reference signals	☹	☺
Effects of speech level in samples	☹	☺
Narrowband (300Hz -3400Hz)	☺	☺
Wideband (100Hz-7000Hz)	☺	Use SWB
Superwideband, SWB (50Hz – 14000Hz)	☹	☺
Linear Frequency distortion sensitivity	☹	☺

The Perceptual Quality Experts.

Appendix B: Reference

B.2 NISQA model architecture

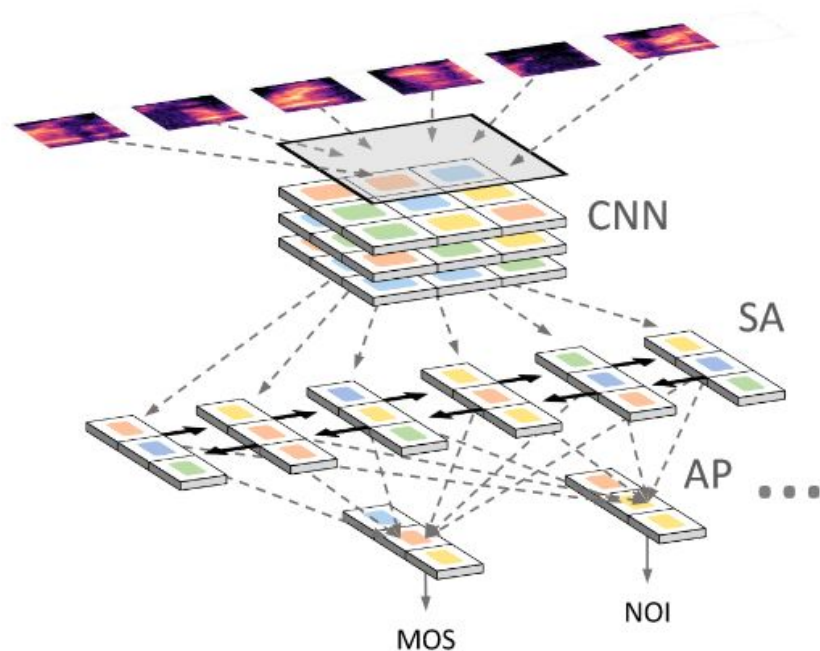


Figure 3: *NISQA neural network architecture.*