Objective perception metrics for speech quality and intelligibility

Natalia Nessler
natalia.nessler@epfl.ch

School of Computer and Communication Sciences

Master Project

March 2021

Paolo Prandoni
LCAV, EPFL

Pablo Mainar
Logitech

Abstract

The evaluation of speech quality is an important task in many development processes. In most of them, it is useful to understand how a *human* would perceive the speech quality. Simulating the human perception algorithmically is a nontrivial task.

In our project, we developed an algorithm evaluating the speech quality objectively and without any reference. Its performance is comparable to the currently used algorithms.

Table of Contents

Introduction

In communication systems, the signal may be affected by background noise, room reverberation and network impairments. Many hardware and software products seek to improve the signal quality — a task for which it is crucial to evaluate the signal quality accurately.

Several techniques are currently used to measure the quality of a speech signal. The most popular ones include PESQ, POLQA, 3QUEST and STOI. These algorithms analyze the signal and output a score specifying the speech quality. Unfortunately, all these techniques are *intrusive*: they require a clean reference signal, while in many use cases only a noisy signal is available, e.g. in a real-time enhancement software. Moreover, it is unclear how these algorithms correlate with human perception of speech quality.

On the other hand, a *non-intrusive* way to evaluate the quality of a signal is the subjective testing. It consists in a survey, where a high number of participants rate the perceived signal quality; a mean opinion score is then computed. This technique is expensive and time consuming, and it is again impossible to evaluate the speech quality in real time.

The goal of this project is to implement an objective and non-intrusive algorithm to evaluate the speech quality with a high correlation to the subjective score. We are going to compare our results to the scores predicted by POLQA, since the latter is widely used as an industrial standard, and in many applications it is required to achieve a minimum POLQA score.

Report organization

In the *State of the art* section, we explore the current solutions, their advantages and drawbacks. We get inspired by some of the ideas to build our model.

The *Data* section describes the data collection. We discuss the techniques used to build the dataset and the organization of the labelling process. We also analyze the resulting subjective scores and compare them to POLQA scores, since POLQA is currently widely used.

In the *Model* section we present the current model and what we have implemented throughout the project. We talk about the issues encountered and the solutions we came up with. We evaluate our results using the mean square error and compare them to POLQA performance, as well as to the variance of the subjective scores.

Finally, we suggest the further improvements in the *Further work* section.

State of the art

There are various techniques used to evaluate the speech quality. Most of them are intrusive: in particular, POLQA [1], PESQ [2], 3QUEST [3] and STOI [4] require a clean reference signal to run a comparison algorithm. Moreover, PESQ and POLQA, despite being widely used in many domains, were developed specifically for distortions introduced by speech compression, thus may not perform well enough in the cases with reverberation and/or background noise.

To avoid the clean reference problem, other methods were developed. The main non-intrusive method is the subjective test, or a survey. ITU-T Recommendation P.835 [5] describes precisely how a subjective test should be designed to collect the answers in an optimal way. This crowd-sourcing method is expensive and time consuming; it is not suitable for a large part of the use cases.

Several non-intrusive algorithms were developed to solve the above issues. The closest to our work is presented in *Non-intrusive speech quality assessment using neural networks* [6], and in the following *Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network* [7]. The authors of this work notice the difference between the objective and the subjective results without discussing the reasons behind it. Their best algorithm performs significantly better than any objective metric, with mean square error of 0.13 on their dataset.

Other works concentrate on predicting the speech intelligibility, like in *Predicting speech intelligibility with deep neural networks* [8]. It is worth noting that the speech intelligibility is only one of the aspects of audio quality.

The paper *WAWEnets: a no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality* [9] presents a new kind of input for the neural network: the raw waveforms are used instead of traditional signal processing feature extraction. Their dataset contains speech in several languages, and the results achieve superior performance. However, this work uses the objective metrics scores as targets for their neural network, which, as shown previously, may not represent well a subjective perception. The same applies to *Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model based on BLSTM* [10].

Data

A large dataset is required in order to implement a meaningful model. It should consist of speech recordings covering a wide range of noise intensity, as well as different types of reverberation. It should also be labeled — we use the subjective testing for this purpose, since it is going to be our target to predict.

Requirements

We start with a clean speech dataset which we will use to mix with room impulse responses and noises. The dataset should consist of clean speech only, with both male and female voices equally distributed. We decide to ignore the children voices for two reasons: they are a lot more difficult to find, and they are not useful for most of our use cases.

Due to the subjective testing, the dataset should have public license, since it will be made available online. To avoid ambiguous ratings, we want the participants to understand the speech; for this reason, we search for English recordings with meaningful sentences, understandable pronunciation, with no foreign accents and no overlapping. Finally, we look for a specific length of recordings: they should consist of one full sentence, long enough for the participants to evaluate its quality, and short enough since the participants must listen to it entirely.

Among numerous speech datasets, the *LibriSpeech* [11] dataset met all our requirements, despite not being recorded in a perfect quiet room. For our purpose, we filter out only the files between 5.5 and 6.5 seconds long. Our resulting clean dataset consists of 236 files sampled with 16 [kHz] rate, where 114 are male recordings and 122 are female.

Mixing

For the reverberation, we use a set of room impulse responses from *OpenSLR28* [12]. It consists of 286 recordings of a short impulse sound in various types of rooms and conditions. The distance to the microphone, the shape and the size of the room are taken into account. For the room impulse responses with several channels, we keep only the first one. To quantify the reverberation, we use *T60*: the time it takes for a sound to decay to 60 [dB].

The computation of T60 is nontrivial. We start with trimming the silence in the beginning of each room impulse response, to avoid undesired artefacts. We then separate each room impulse response in bands of frequencies by applying a bandpass filter; for each band, we normalize the filtered signal and compute the Schroeder integral. A linear regression is then applied to estimate the slope between -5 [dB] and -35 [dB]. This way we obtain an estimation of T30; T60 is obtained by multiplying T30 by 2. Finally, we keep the median of all the bands per room impulse response as our value of the reverberation time.

For the background noise, we use the *FreeSound* [13] dataset. We choose only 9 types of noise that fit our use cases the best: typing, squeaky chair, air conditioner, copy machine, shutting door, cafeteria, restaurant, babble and traffic.

Both reverberation and noise files are sampled with rate 16 [kHz]. We mix the clean speech samples with the room impulse responses and noises in the following way (Figure 1):

1. Each clean sample is convolved with three room impulse responses picked randomly without replacement. The "convolutional tail" is trimmed.

2. For each of the resulting samples, three different kinds of noise are picked randomly with a signal to noise ratio chosen uniformly at random between -5 and 25 [dB]. Since the noise files are much longer than the clean samples, a random interval is chosen to be added.

3. The background noise is also added once to the original clean sample, using the same method.

4. Finally, the three convolved samples without noise are also included in the final dataset, as well as the original clean sample.

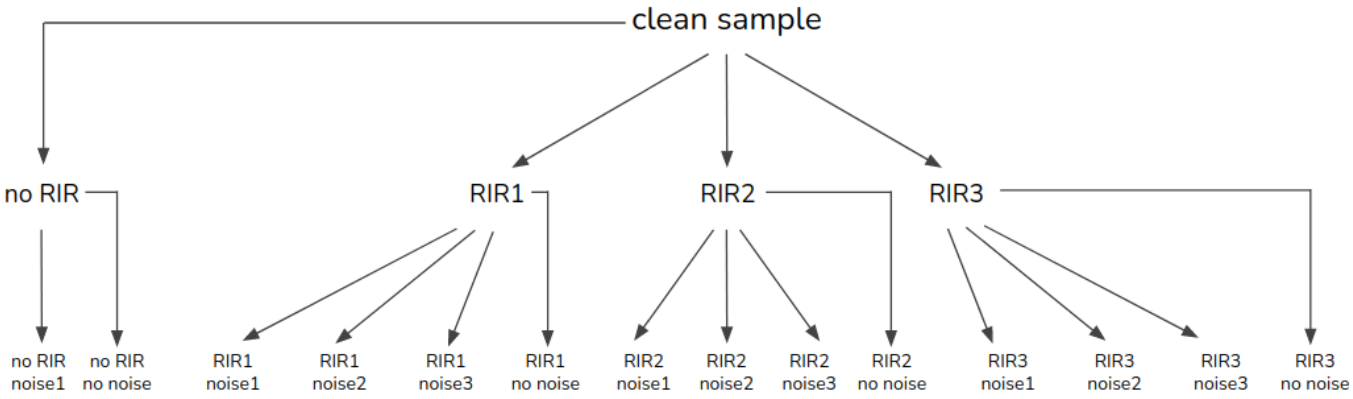

Figure 1. Mixing scheme. RIR stands for room impulse response.

The resulting dataset consists of 4032 samples with total length of 6.7 hours. Figure 2 shows the distribution of T60 and SNR in the dataset.

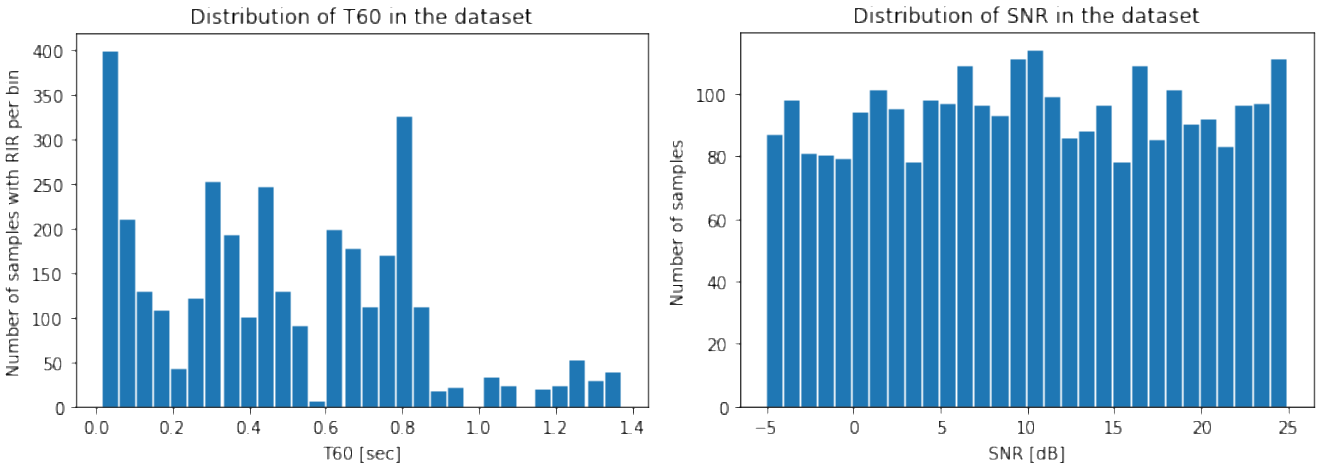

Figure 2. Distribution of T60 and SNR in the mixed dataset.

Subjective testing

In order to train the machine learning model, we need a labeled dataset. For this purpose, we use Amazon Mechanical Turk [14]: an online platform allowing to run surveys and pay the workers for their answers. It has many features allowing to easily review, accept or reject the answers, as well as submit the payments automatically.

Due to the important size of the dataset, we must divide it into multiple smaller tests; this implies publishing multiple surveys on Amazon Mechanical Turk. To generate them automatically, we use the code provided by Microsoft [15], [16]: it follows the ITU-T Recommendation Standard P.835 [5] which describes in detail how exactly a subjective test evaluating speech quality should be designed.

The subjective test is divided in *HITs*: Human Intelligence Tasks. A HIT consists of multiple questions that are answered by a worker in a limited amount of time and then submitted. The worker can continue working on the next HIT and is paid by HIT. In our case, a HIT consists of four sections:

Qualification

In this section, the worker answers several general questions to verify the eligibility. In particular, the worker should be English native speaker and use in-ear or over-ear headphones. We also verify the hearing ability: five audio files are given where a male voice says three digits with noise in the background; the worker must write down the digits.

This section is shown only once per worker; once it is completed, the next HITs skip it automatically for this worker.

Setup

Here, the worker adjusts the computer volume and is asked to not change the volume afterwards until the end of the test. Then, the worker is asked to play a file where several digits are said and write them down; some of the digits are only given in the left ear, and the other ones are given in the right ear. This allows to verify that the worker uses a stereo headset and hears well with both ears.

Finally, the worker is asked to compare the quality of four pairs of speech samples 6 seconds long. Each pair consists of the same sample with the same background noise at two different levels: with SNR equal to 40 [dB] in one case and 50 [dB] in the other one. The worker is asked to choose the sample with best quality in each pair or say that the difference is not detectable. This ensures that the worker can recognize the difference of 10 [dB] in the SNR level and therefore doesn't use any noise-cancelling device which could interfere with the results of our test.

The setup section reappears every 60 minutes, implying that a worker can submit several HITs before going through this section again.

Training

This section is similar to the real test, but the answers of the workers are not taken into account; it is only here to prepare the worker for the real test. It consists of 4 speech samples from our dataset. These samples should cover the entire range of quality, in order to show to the worker the various conditions present in the dataset.

For each of the samples, the worker must answer to the following questions by choosing one of the options:

1. Attending only to the **speech signal**, select the category which best describes the sample you just heard. The **speech signal** in this sample was:
    - Not distorted — 5
    - Slightly distorted — 4
    - Somewhat distorted — 3
    - Fairly distorted — 2
    - Very distorted — 1
2. Attending only to the **background**, select the category which best describes the sample you just heard. The **background** in this sample was:
    - Not noticeable — 5
    - Slightly noticeable — 4
    - Noticeable but not intrusive — 3
    - Somewhat intrusive — 2
    - Very intrusive — 1
3. Select the category which best describes the sample you just heard for purposes of everyday speech communication. The **overall speech sample** was:
    - Excellent — 5
    - Good — 4
    - Fair — 3
    - Poor — 2
    - Bad — 1

For each of the three questions, the worker must listen to the sample first. Only after the sample is played until the end, the answers are made available for selection.

To avoid bias, the order of the first two questions is chosen randomly for each HIT and is kept constant for the entire HIT, so that the worker can get used to the order.

This section appears every 8 hours: if the worker continues working on the test during the day, there is no need to remind how the questions are, but it may be useful if the worker returns on the next day.

Ratings

This is the last section containing the real test. It looks exactly like the Training section, but consists of 30 samples to rate instead of 4. The order of the samples in the HIT is random for each worker. Among these samples, 28 are the samples from our dataset. In addition, there are one *gold question* and one *trapping stimulus*.

The gold question is used as a hidden quality control item. It is a sample from our dataset for which we expect a particular answer to the overall quality question. We choose these samples manually to verify that the answer is indeed obvious for all participants. The samples have either very good quality (clean speech without any reverberation or noise) or very bad quality (high T60 and negative SNR). The worker passes this quality check if the given answer is at most 1 score away from the expected answer: 4 or 5 for the good quality samples and 1 or 2 for the bad ones.

The trapping stimuli are used to motivate the workers to answer the questions attentively, as well as to check that they are paying attention. A trapping stimulus is an audio file that begins like any other sample from the dataset; after 2 seconds, the usual sample is interrupted with a motivational message: *"This is an interruption. Please select the answer X to confirm your attention now"*. The worker passes this quality check if the answer to all three questions corresponds to the score given in the message. It has been shown in [17] that the presence of the trapping stimuli in the subjective test improves the quality of the answers.

In total, 144 HITs are created. We add three qualification requirements on the worker: the number of approved HITs should be greater than 500, the HIT approval rate should be greater than 98%, and the worker should be located in Australia, Canada, Ireland, New Zealand, Great Britain or the United States, since there is no language requirement available on the platform. Only the workers matching these requirements can see our test.

We require 9 submissions per HIT, after which the HIT is automatically hidden from workers. This number is chosen based on the platform fees which double from 20% to 40% starting from 10 submissions per HIT.

Each worker can submit at most 30 HITs to avoid bias introduced by too many samples rated by the same worker. The workers are paid according to the minimum wage in the United States, since most of the workers are expected to be located there, and we estimate the time required to submit one HIT around 30 minutes.

Acceptance criteria

Once all the answers are collected, we download the results and run a script to analyze the answers. We reject the entire submission if at least one of the following conditions is true:

1. The digits in the setup section are wrong, implying that the worker hears only one of the channels.

2. The answer to the trapping stimulus is wrong, implying that the worker hasn't been attentive enough.

3. The answer to the gold question with bad quality sample is greater or equal to 3.

4. The variance of the answers through the HIT is below 0.2. Given that there is one trapping stimulus with a fixed answer, a variance below 0.2 implies almost always the same score for all the other questions.

The workers are not paid for the rejected submissions. Some other submissions are not good enough to be used in our data, but their workers still deserve being paid for their work. This is the case when at least one of the following conditions is true:

1. The answer to the gold question with good quality sample is less or equal to 3. In fact, since the *LibriSpeech* dataset has been recorded by volunteers in usual quiet conditions, but not in a perfect quiet room, even a clean sample could be considered by some workers as medium quality.

2. Among the four pairs of samples in the setup section, at most one has been chosen correctly.

The rest of the answers are accepted and used in our data, leading to the following values: 287 workers submitted HITs with an average of 106 samples per worker. There are 7.5 scores per sample in average; Figure 3 shows the exact numbers. Among the few samples with 4 or 5 ratings only, those with the standard deviation above 0.5 were discarded. In other words, we keep the samples where the difference between the lowest and the highest ratings does not exceed 1. For now, only the answers to the overall quality question are used.

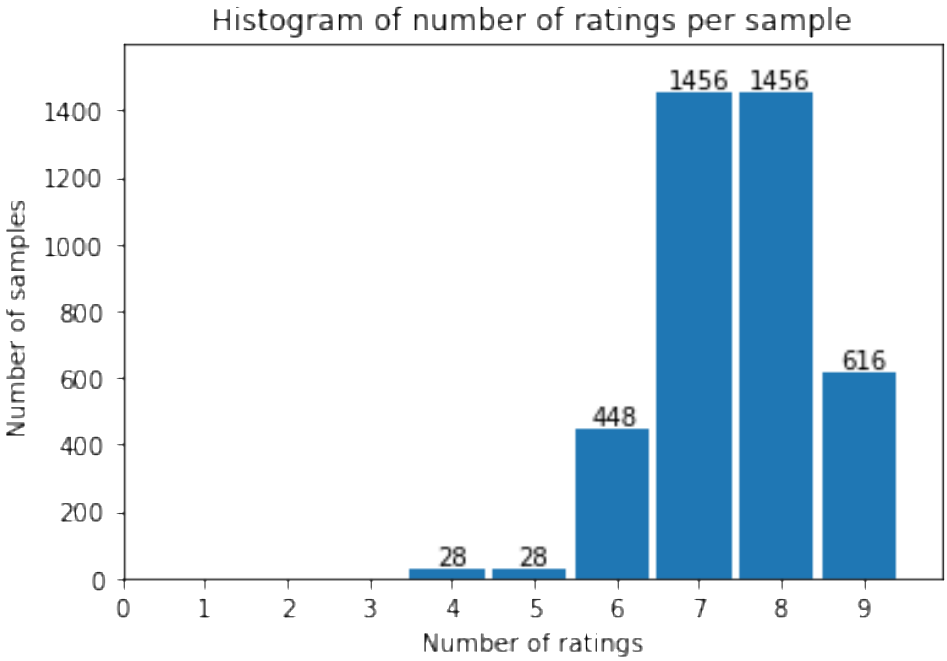

Figure 3. Histogram of the number of ratings per sample.

Scaling

The mean opinion score and the standard deviation are computed for each sample. We observe that some workers don't use the entire range of scores available. Figure 4 shows that, for example, 47 workers rated the samples from 1 to 4 or from 2 to 5. This implies that a medium score is ambiguous depending on the worker.

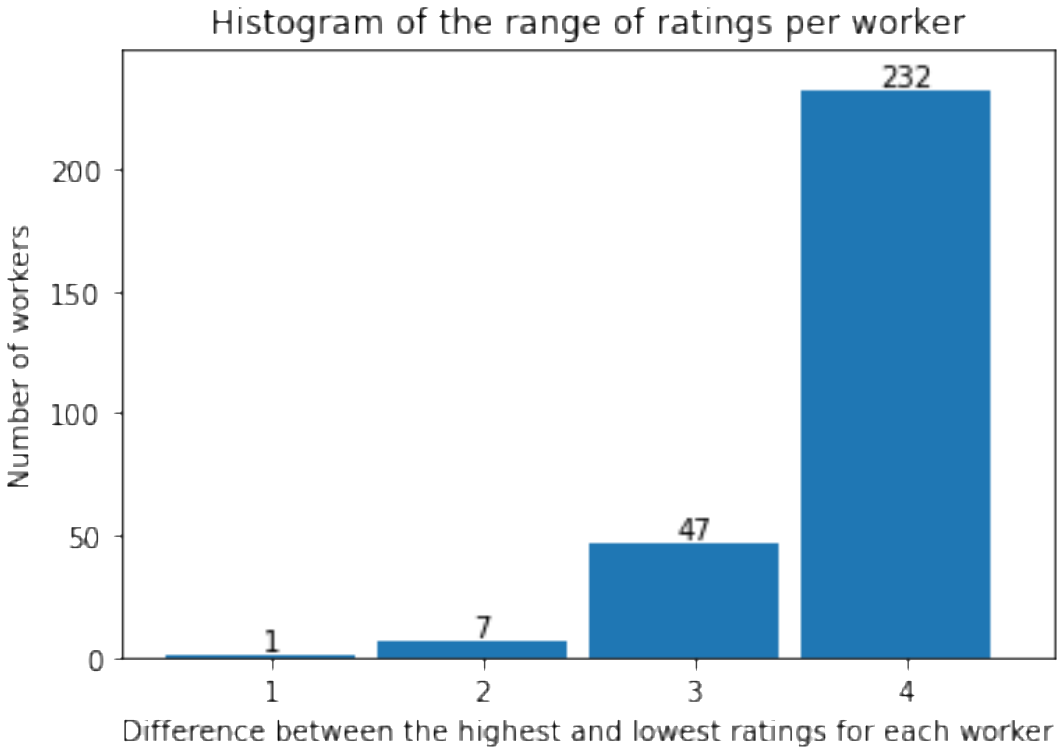

Figure 4. Histogram of the range of ratings per worker.
For example, 232 workers used the entire range from 1 to 5.

For this reason, we decide to scale the ratings of all workers and bring them to the general range of ratings between 1 and 5. An exception is made for the worker who used only the scores 3 and 4: we scale these ratings to the range between 2 and 4, because the method above would let them be only either 1 or 5, which is too extreme. The scaled ratings are only used in our last model, since we used this technique towards the end of the project.

Analysis

Before implementing the algorithm, it is interesting to visualize the data. Figure 5 shows the values of the mean opinion score and the standard deviation plotted against T60 and SNR. Each dot represents a sample, and its color corresponds to the value of the statistic. We observe that the scores correlate well with both T60 and SNR. The samples with no noise are rated in general higher despite the reverberation — the MOS decays slowly in this case. The workers have been more sensitive to the presence of the background noise even in the case when there is no reverberation. The standard deviation does not change much depending on T60 or SNR, as expected. Workers tend to agree better when SNR is low.

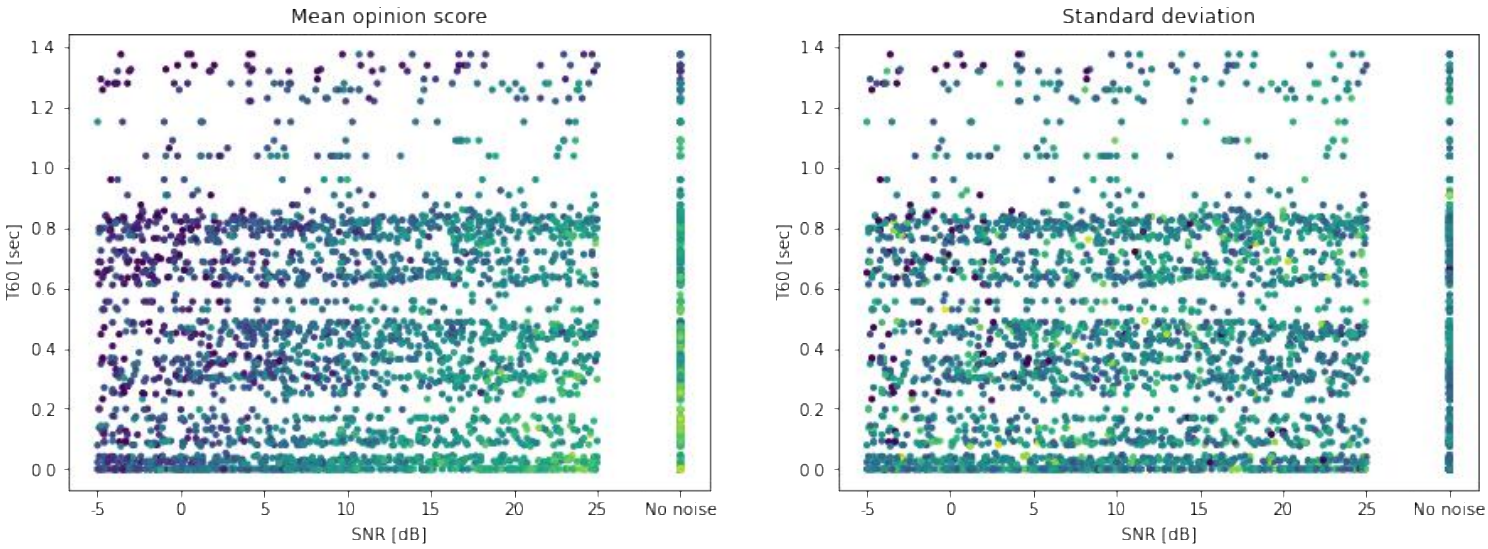

Figure 5. Mean opinion score and standard deviation with respect to T60 and SNR. Bright color means high value, dark color — low value.

The type of noise seems to influence the mean opinion score slightly. According to Figure 6, the worst noises are uniform over time, except for the air conditioner. Since the SNR is distributed uniformly over the dataset, it has no impact on the mean opinion scores here.

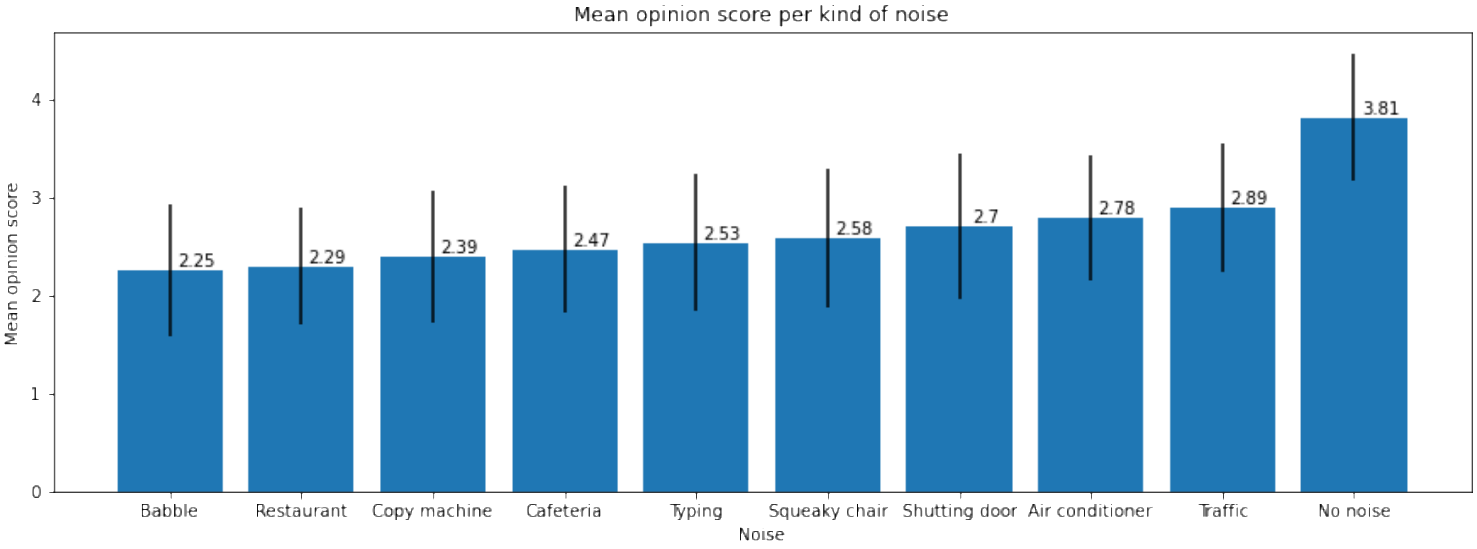

Figure 6. Mean opinion score per type of noise. SNR is uniform over the dataset and has no impact on the values.

Comparison between the subjective test and POLQA

It is interesting to verify how POLQA correlates with the subjective ratings. Figure 7 shows the histograms of the scores: we observe that the shapes are very different — POLQA tends to rate lower.

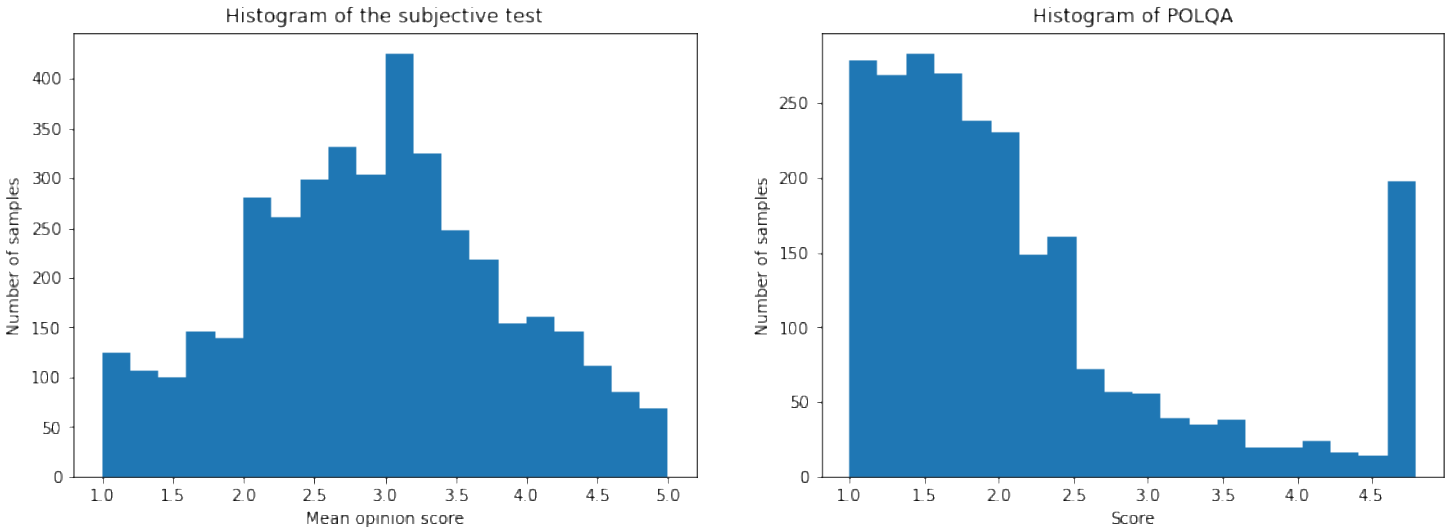

Figure 7. Comparison between the subjective test (left) and POLQA (right).
The right column in the POLQA histogram is due to the samples without any noise or reverberation: POLQA outputs a constant maximum value if the reference signal is the same as the evaluated one.

The reason behind the above difference is shown on Figure 8: while some samples get similar ratings in both systems, other ones get a high subjective score and a low POLQA score. We don't see any reciprocal cases.

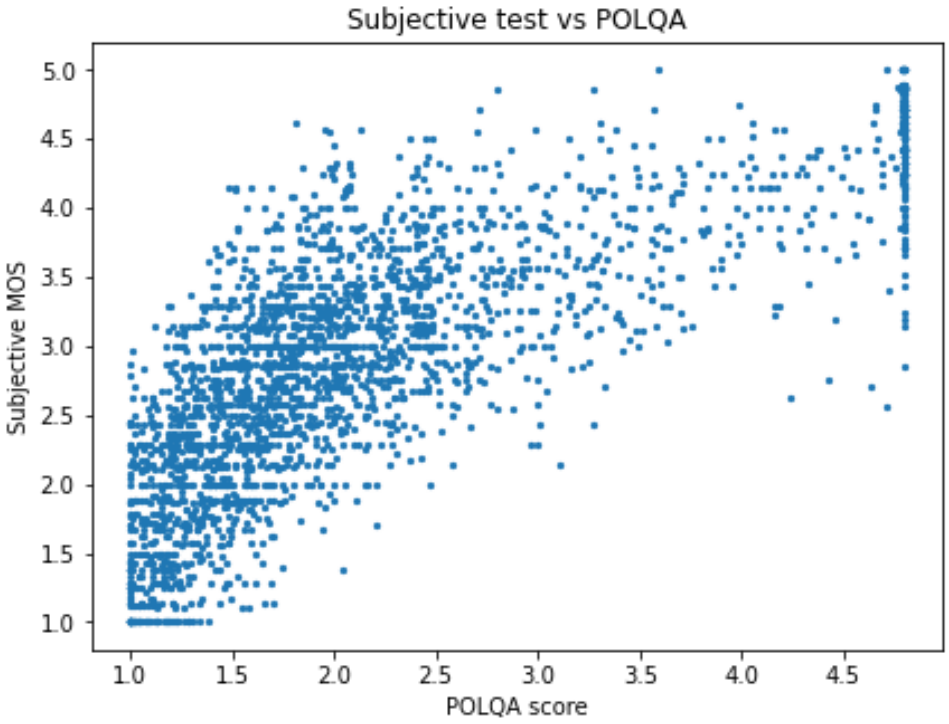

Figure 8. Comparison of the subjective scores to POLQA scores

We want to find the samples for which the scores of POLQA and the subjective test differ the most. According to the data, the vast majority of such samples have no noise, but have reverberation. The rare samples with noise have SNR above 20 and the background noise

among those on the right side of Figure 6. We observe this on Figure 9: the bottom line of the plots has similar values — these samples have no reverberation. As soon as T60 is increased, POLQA scores decrease very fast, while the subjective scores remain high for medium to high SNR values. The samples without noise show the greatest difference between the subjective scores and POLQA.

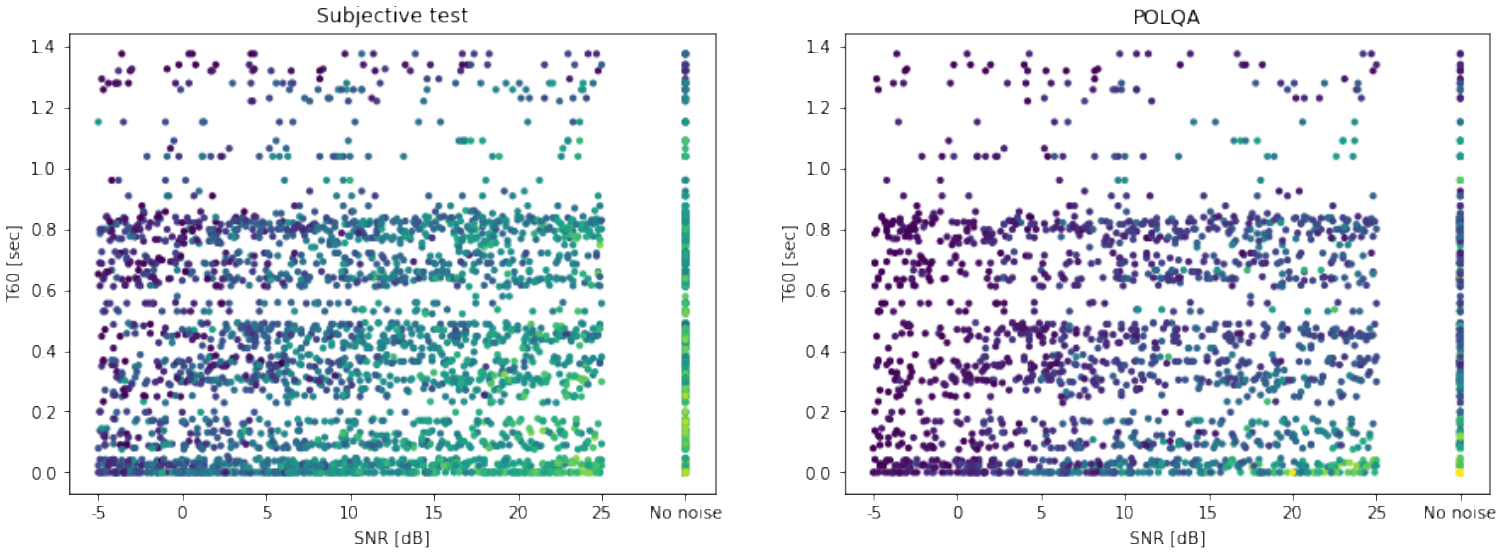

Figure 9. Comparison of the MOS between the subjective test (left) and POLQA (right). Bright color represents high value, dark color — low value.

This can be explained by the fact that a human is used to some reverberation — not only in the communication systems, but even in the real life. A reverberant speech without noise is intelligible. As soon as the background noise is added, the subjective perception becomes less tolerant to the impairment. On the other hand, POLQA has not been intended for evaluating reverberant speech: according to *Recommendation P.863.1* [18], T60 should be below 0.3 [sec] above 200 [Hz], with the distance to the microphone of approximately 10 [cm]. Nevertheless, POLQA is widely used as an industrial standard, and in many applications it is required to achieve a minimum POLQA score. In these cases, an alternative objective speech metric may become useful.

Model

To make the speech samples understandable by our model, we must extract important features from them.

Features

The most common features are *Mel-frequency cepstral coefficients* (MFCC). They are obtained for each sample as follows [19]:

1. A Fourier transform of the signal is computed.
2. The powers of the spectrum are mapped on the *mel scale*[1] using overlapping 13 triangular windows.
3. A base 10 logarithm of the powers is computed for each mel frequency.
4. A discrete cosine transform (type II) is applied to the obtained "signal" to reduce dimensionality.
5. The amplitudes of the spectrum are the Mel-frequency cepstral coefficients.

We normalize the MFCC and obtain a 13×172 matrix for each sample. Figure 10 shows an example of two extreme cases with the best quality on the left and the worst quality on the right. The lowest row is the DC component and can be ignored.

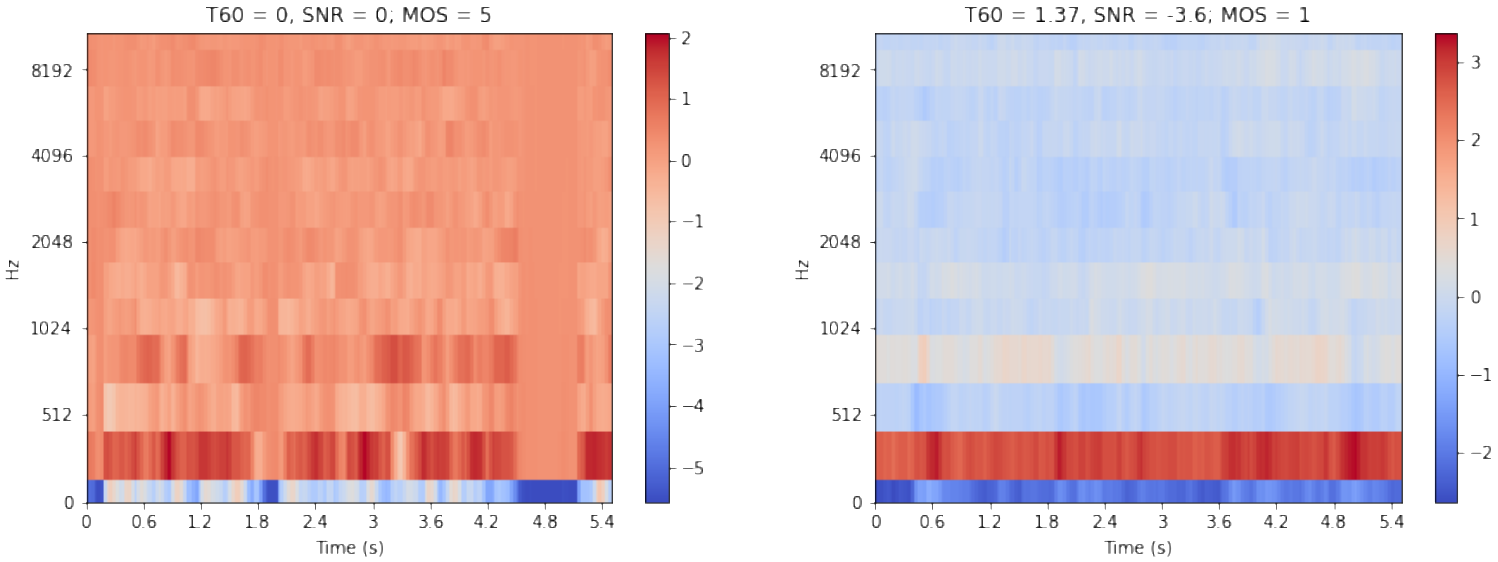

Figure 10. An example of MFCC in two extreme cases.
On the left: a sample with no reverberation and no noise, with the subjective score 5.
On the right: a sample with the highest T60 in the dataset and very low SNR, labeled 1.
The labels on the y-axis are frequencies instead of mels, because the exact definition of mel is ambiguous.

We split the shuffled data: our training set consists of 80% of samples, with 20% kept for the test set.

[1] The mel scale is a scale of pitches, where the distance between any two consecutive pitches is perceived by a listener equally. 1000 [mels] correspond to 1000 [Hz].

Baseline

We start by implementing a simple model as our baseline. It has the following shape:

Layer	Parameters	Output shape
Convolutional layer	Filters: 5, kernel size: 2×2	12 × 171 × 5
Max-pooling	Kernel size: 1×3, strides: (1, 2)	12 × 85 × 5
Convolutional layer	Filters: 5, kernel size: 2×2	11 × 84 × 5
Flatten	—	4260 × 1
Fully connected layer	Units: 1	1 × 1

Through the entire network we use the same rectified linear activation function (ReLU). We use Adam optimizer and measure the loss using the mean square error value.

After 35 epochs, the loss converges to 0.66 over our test set. It is already much lower than the POLQA predictions — the latter show a MSE of 0.99 on the same test set.

Complex model

We can now move to a more complex model. We start by trying a model similar to the one presented in *Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network* [7]. It takes as input the list of MFCC, where each element has a shape of 13 × 172, and consists of the following layers:

Layer	Parameters	Output shape
Convolutional layer	Filters: 16, kernel size: 2×2	12 × 171 × 16
Batch normalization	—	12 × 171 × 16
Max-pooling	Kernel size: 1×3, strides: (1, 2)	12 × 85 × 16
Convolutional layer	Filters: 32, kernel size: 2×2	11 × 84 × 32
Batch normalization	—	11 × 84 × 32
Max-pooling	Kernel size: 3×3, strides: (2, 2)	5 × 41 × 32
Convolutional layer	Filters: 64, kernel size: 2×2	4 × 40 × 64
Batch normalization	—	4 × 40 × 64
Convolutional layer	Filters: 32, kernel size: 2×2	3 × 39 × 32
Batch normalization	—	3 × 39 × 32
Flatten	—	3744 × 1
Fully connected layer	Units: 128	128 × 1
Dropout	Rate: 0.5	128 × 1
Fully connected layer	Units: 128	128 × 1
Dropout	Rate: 0.5	128 × 1
Fully connected layer	Units: 1	1 × 1

As before, the activation function is ReLU, we use Adam optimizer and evaluate the loss as mean square error. In total, 515217 parameters are used.

This model performs significantly better than the baseline. The MSE over our test set is equal to 0.3 after 40 epochs.

Pre-training on POLQA scores

In every machine learning problem, a bigger dataset implies better results. However, we cannot easily increase the size of our dataset, since the labeling of the speech samples is expensive. Instead, we can pre-train our model on another dataset — labeled by POLQA. This way, we can create a much larger dataset and train our model to predict the POLQA scores first; after that, we can initialize our main model with the output of the previous training, which is expected to be more accurate than randomly generated values.

We generate a new dataset following the same steps as before. It consists of 25360 samples, which is 6 times more than in our previous dataset. The overall length of this dataset is 42.2 hours of speech. The new dataset is labeled by POLQA. We feed it to our model with POLQA scores as a target; after 50 epochs, we achieve a MSE of 0.33.

Now, we keep the same model and train it on top with our previous dataset and subjective scores as a target. Again after 50 epochs, we get a MSE of 0.17, which is significantly lower than our previous results on the small dataset.

MOS scaling

Finally, we use the mean opinion score scaling discussed in the *Scaling* section. We expect the dataset quality to be improved. We reuse the model pre-trained on POLQA and train it again with the scaled subjective dataset. This method shows the best results with MSE equal to 0.11 after 5 epochs. With more epochs, the model overfits and the validation loss starts to grow, which can be observed on Figure 11.

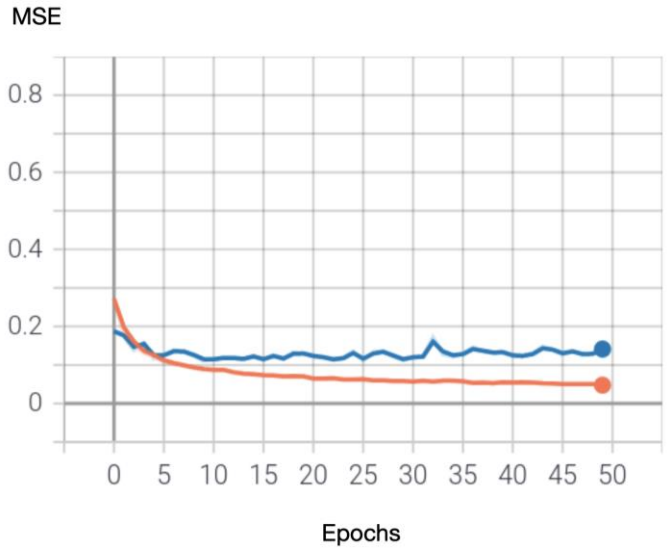

Figure 11. Evolution of the training and validation losses. Orange line: training loss, blue line: validation loss.

Discussion

We tried several models, consistently improving our results. It is interesting to compare the methods:

Method	**MSE** (with respect to the subjective test)
POLQA	0.99
Baseline	0.66
Complex model	0.3
Complex model pre-trained on POLQA scores	0.17
Complex model, pre-trained, with MOS scaled	**0.11**

We observe that even a simple low-complexity model like our baseline performs much better than POLQA. When we increase the complexity of the model, the MSE decreases significantly. The pre-training on POLQA scores shows good results as well, making use of a larger dataset. Finally, the MOS scaling improves our results even more.

It is interesting to compare our MSE to the variance of the MOS through our dataset. In fact, the mean variance of the MOS is equal to 0.56 in the original dataset and to 0.5 in the dataset with scaled MOS. This implies that on average, the listeners themselves may rate a sample further away from the mean score than our model.

The complexity of the model does not affect its usability once it is trained. It takes on average 0.1 seconds to evaluate the quality of one sample (6 seconds long) with the trained model[2].

[2] With the processor: Intel Core i5-7300U CPU @ 2.60GHz × 4

Further work

The work on this project being limited in time, there are still a lot of improvements to try.

Data

As mentioned at the end of the previous section, it is difficult to evaluate the speech quality based on the subjective test better than the listeners do it themselves. To solve this issue, more analysis can be done on the data. We have shown that the MOS scaling improves the results, meaning that we understood correctly the implied quality behind the actual scores. It might be helpful to find more such improvements to do on the dataset.

Moreover, we have not used all the data available. As explained in the *Subjective testing* section, listeners were asked to rate every sample three times: based on the speech quality, noise intrusion and overall impression. For now, we used only the overall scores; the two other scales may provide us with some insight on the listener's perception.

Finally, the speech samples themselves could also be improved: the dataset may lack for perfectly clean samples, since the LibriSpeech dataset has been recorded in natural environment. Moreover, the room impulse responses should also be adapted. Figure 2 shows the distribution of T60 in our dataset: in contrast to SNR, it is far from being uniform. This may introduce some bias in our dataset according to *Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation* [20]. Ideally, both T60 and DRR[3] (which has not been computed for our dataset) should be uniform over their entire range.

Various features

For now, only MFCC features have been used in our model. Adding more different features can help with extracting non-obvious information from the samples. We find at least the following features interesting:

- Pitch
- Voice activity
- Frame energy
- Raw waveforms [9]

[3] Direct to reverberant ratio

Transfer learning

Pre-training on a larger dataset with POLQA scores as target worked well for us. However, we could go further and use an already pre-trained audio-related model to extract embeddings for our own model. This is similar to the idea above of adding various features, but in a more complex way, since a pre-trained model may extract some hidden information we are not aware of.

Network impairments

In our project, we focused on two main impairments: the reverberation and the background noise. These are not the only impairments that may affect the speech quality in communication systems. In particular, the packet loss and jitter happen very often and must be handled separately. It would be interesting to at least test the performance of our model in such cases, and ideally train our model on a more exhaustive dataset.

Conclusion

Throughout our project, we studied various techniques of evaluating the speech quality. We used three of them: the subjective testing, POLQA and our own model. For this purpose, we have generated a large dataset consisting of speech samples with different kinds of reverberation and background noise. A small part of this dataset has been labeled subjectively, while the rest of it has been evaluated by POLQA.

We compared the scores of the subjective test to the output of POLQA and observed that POLQA does not perform well in predicting the subjective perception when reverberation is present. Nevertheless, POLQA is widely used as an industrial standard, and in many applications it is required to achieve a minimum POLQA score. For this reason, an alternative objective speech metric may become useful.

We have implemented a machine learning model and pre-trained it to predict the POLQA scores first; we then have trained it again with the subjective scores as our target. Combined with the subjective scores scaling, our model achieves a mean square error of 0.11 on our dataset, which is lower than the variance of the listeners' scores per sample. Apart from showing better results than the currently used techniques, our model is both objective and non-intrusive, implying that its usage is fast, not expensive and does not require any clean reference signal.

References

[1] ITU-T, “Recommendation P.863: Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals”, January 2011.

[2] ITU-T, “Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for endto-end speech quality assessment of narrowband telephone networks and speech codecs”, February 2001.

[3] HEAD acoustics, “3QUEST: 3-fold Quality Evaluation of Speech in Telecommunications”, October 2008.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech”, IEEE Trans. ASLP, vol. 19, no. 7, pp. 2125–2136, September 2011.

[5] ITU-T, “Recommendation P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm”, November 2003.

[6] Anderson R. Avila, Hannes Gamper, Chandan Reddy, Ross Cutler, Ivan Tashev, Johannes Gehrke, “Non-intrusive speech quality assessment using neural networks”, March 2019.

[7] Hannes Gamper, Chandan K A Reddy, Ross Cutler, Ivan J. Tashev, Johannes Gehrke, “Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network”, October 2019.

[8] Constantin Spille, Stephan D. Ewert, Birger Kollmeier, Bernd T. Meyer, “Predicting speech intelligibility with deep neural networks”, Computer Speech & Language, vol. 48, pp. 51–66, October 2017.

[9] Andrew A. Catellier, Stephen D. Voran, “WAWEnets: a no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality”, May 2020.

[10] Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, Hsin-Min Wang, “Quality-Net: An End-to-End Non-intrusive Speech Quality Assessment Model based on BLSTM”, August 2018.

[11] Vassil Panayotov, Guoguo Chen, Daniel Povey, Sanjeev Khudanpur, “LibriSpeech: an ASR corpus based on public domain audio books”, April 2015.

[12] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael LSeltzer, Sanjeev Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” IEEE ICASSP, March 2017.

[13] Reddy Chandan KA, Beyrami Ebrahim, Pool Jamie, Cutler Ross, Srinivasan Sriram, Gehrke Johannes, “A Scalable Noisy Speech Dataset and Online Subjective Test Framework”, Proc. Interspeech 2019, pp. 1816-1820, 2019.

[14] Crowston K., “Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars”. In: Bhattacherjee A., Fitzgerald B. (eds) Shaping the Future of ICT Research. Methods and Approaches. IFIP Advances in Information and Communication Technology, vol 389. Springer, Berlin, Heidelberg, 2012.

[15] P.808 Toolkit, website.

[16] Babak Naderi, Ross Cutler, “An Open Source Implementation of ITU-T Recommendation P.808 with Validation”, Proc. Interspeech, May 2020.

[17] Babak Naderi, Tim Polzehl, Ina Wechsung, Friedemann Köster, Sebastian Möller, “Effect of Trapping Questions on the Reliability of Speech Quality Judgements in a Crowdsourcing Paradigm”, September 2015. P.863.1 : Application guide for Recommendation ITU-T P.863

[18] ITU-T, “Recommendation P.863.1: Application guide for Recommendation ITU-T P.863”, June 2019.

[19] Sahidullah Md., Saha Goutam, “Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition”, May 2012.

[20] Nicholas J. Bryan, “Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation”, May 2020.