Performance of Objective Speech Quality Metrics on Languages Beyond Validation Data: A Study of Turkish and Korean

Javier Pérez López[1]

Supervisor(s): Jorge Martínez Castañeda[1]**, Dimme de Groot**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2025

Name of the student: Javier Pérez López
Final project course: CSE3000 Research Project
Thesis committee: Jorge Martínez Castañeda, Dimme de Groot, Przemyslaw Pawelczak

Abstract

This study investigates the performance of two objective speech quality metrics, Perceptual Evaluation of Speech Quality (PESQ) and Virtual Speech Quality Objective Listener (ViSQOL), in predicting human-rated speech quality scores, which are essential for telecommunication systems' Quality of Experience (QoE). These metrics have been validated using a limited number of languages due to the insufficiency of labeled data with human-rated scores. This research focuses on the applicability of PESQ and ViSQOL in Turkish and Korean, two languages that were not part of the validation data for calibrating these metrics.

The experiment used English as the baseline language for comparison, and the results showed that Turkish samples had higher average ViSQOL scores, with the difference being statistically significant compared to the English samples. Furthermore, Turkish male speakers had the highest correlation between PESQ and ViSQOL scores, and ViSQOL rated speech higher than PESQ, especially under babble noise degradations. Future research should focus on extending this study by exploring biases across additional metrics and languages, while also constructing a dataset with labeled subjective scores for more languages to improve the calibration of these metrics.

1 Introduction

Measuring speech quality in telecommunication systems is essential to ensure optimal user experiences, and has been a topic of extensive research driven by the need to adapt to evolving technologies. Perceptual audio quality refers to the way sound is perceived by human listeners. To evaluate this quality, the International Telecommunication Union (ITU-T) has established a range of standardized metrics over time, incorporating both subjective and objective methods for assessing speech quality in telecommunication systems [1] [2] [3].

Subjective methods involve human participants rating the perceived quality of speech signals on a predefined scale. The most commonly used approach is the Mean Opinion Score (MOS), standardized in ITU-T Recommendation P.800 [4], in which, participants assign scores ranging from 1 ("bad") to 5 ("excellent"). The MOS is calculated as the mean of these individual ratings.

While accurate, these methods are time-consuming and costly. In contrast, objective metrics provide an automated approach to predicting speech quality that aims to emulate human ratings, offering a faster, scalable, and more efficient alternative [3].

For this reason, objective speech quality metrics have been a significant research focus, leading to the standardization of PESQ by ITU-T in 2000 [5], and its successor, Perceptual Objective Listening Quality Analysis (POLQA) [6], in 2011. Over time, additional metrics such as ViSQOL [7], Non-Intrusive Speech Quality Assessment (NISQA) [8], and 3-fold Quality Evaluation of Speech in Telecommunications (3QUEST) [9] have been developed.

These objective metrics analyze audio signals to simulate the human auditory system, producing scores that are then mapped to subjective MOS values. The performance of these metrics is then measured by the correlation between objective quality scores and their corresponding subjective MOS scores [10]. Due to the limited availability of human-rated speech quality data, these mappings have been developed and validated using data from a restricted set of languages. Specifically, PESQ's mapping is based on only nine languages [11], while POLQA's mapping is based on ten languages [6]. This approach assumes that the linguistic features present in the training data are sufficient to generalize across different languages, dialects, and accents.

As a result, speech quality assessments for languages outside their original validation data may be affected. Given the rapid expansion of multilingual speech processing applications and telecommunication systems used across the globe, growing concerns have been raised about the applicability of these metrics to languages outside their validation set [2]. Therefore, understanding the robustness of these metrics across diverse languages is critical to ensure accurate quality assessments.

To address these concerns, this study investigates the performance of two popular objective metrics. Specifically, it focuses on PESQ and ViSQOL. A more in-depth explanation of these metrics can be found in section 2. This study evaluates their performance in two languages, Turkish and Korean, and compares the results to those of English, which is used as the reference for comparison. Turkish is spoken by approximately 75 million people and accounts for 1.8% of content on relevant websites across the Internet, with an internet penetration rate of 86.5% as of early 2024 [12] [13] [14]. Korean, spoken by around 80 million people, makes up 0.8% of such content, with an internet penetration rate in South Korea of 97.2% [15] [13] [16]. Due to these factors, both languages have been considered relevant for research purposes.

The main research question is the following:

- How does the performance of PESQ and ViSQOL vary in predicting speech quality for Turkish and Korean, two languages outside their mapping function validation set, considering the effects of gender and different degradation types?

Limited research has been conducted in this area. Konane et al. [17] examined the performance of the PESQ and POLQA metrics for speech quality in two local languages of Burkina Faso (Moore and Dioula), comparing them to English and French. They concluded that, while Moore and Dioula influence PESQ, they do not appear to affect POLQA.

Additional research, such as that by Ben Ali et al. [2], investigated the performance of PESQ for Arabic speech. Their study showed that Arabic speech contains more stationary regions than English, which may lead to PESQ scores being more accurate for Arabic speech under similar network conditions.

This study aims to extend previous research by examining the generalizability of PESQ and ViSQOL when applied to languages outside their mapping function validation set.

The rest of the paper proceeds as follows: in section 2, a brief introduction about objective speech quality metrics is given. Section 3 covers the methodology used to gather, process the data, and perform the experiment. Section 4 discusses the results extracted from the experiment. Section 5 covers the responsible research practices and reproducibility of the work. Lastly, section 6 summarizes the research conclusions and discusses recommendations for future work.

2 Overview of Objective Speech Quality Metrics

Objective speech quality metrics can be categorized into two types:

- **Intrusive Metrics**: Compare a degraded speech signal to a clean reference signal to estimate quality. Some examples include PESQ, POLQA, and ViSQOL. These models calculate a perceptual distance between the reference and degraded signal to approximate how a human listener might rate the quality [18]. They are also referred to as full-reference metrics [19].
- **Non-Intrusive Metrics**: Evaluate the quality of speech without the need for a reference signal. They analyze characteristics of the degraded signal alone by comparing it to an estimation of the reference signal, or checking the degraded signal to find unnatural patterns [3]. They are also known as single-reference metrics. Examples include ITU-T P.563 [20] and more modern deep learning-based approaches.

2.1 Perceptual Evaluation of Speech Quality (PESQ)

PESQ is an intrusive objective speech quality metric that was standardized by the ITU-T as Recommendation P.862 in 2001 [5] [21]. It was developed to predict the subjective quality of narrowband telephony systems and speech codecs designed for narrowband communication (300-3,400 Hz) [5].

PESQ operates as an intrusive metric, requiring both the reference and degraded signals for analysis. The algorithm consists of several steps. First, it aligns the reference and degraded signals to the same power level as the one used for subjective tests. It then compensates for filtering effects from the network, adjusts for any timing discrepancies (such as variable delays), and simulates the signal through the human auditory system. Finally, it calculates disturbance parameters, including symmetric and asymmetric disturbances, to quantify the perceived audio quality [22]. An overview of these processing steps is shown in Figure 1.

PESQ produces raw scores in the range of -0.5 to 4.5, which are not directly comparable to subjective MOS scores. To address this, ITU-T Recommendation P.862.1 introduced a third-order polynomial mapping function that transforms the raw scores into MOS Listening Quality Objective (MOS-LQO) values, allowing for consistent comparisons with other quality metrics. The resulting MOS-LQO scores for speech are in the range of 1 to 5 and were optimized using subjective data across various applications and languages. Specifically, the following nine languages were considered: British English, American English, Swedish, French, Italian, German, Finnish, Dutch, and Japanese [11]. This mapping has a strong accuracy in predicting MOS-LQO scores, with results indicating that 93.5% of its predictions fall within 0.5 MOS of the correct subjective scores. However, these tests were conducted only in the languages previously mentioned.

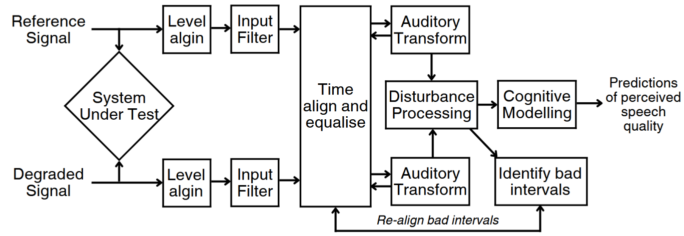

Figure 1: Diagram illustrating key steps in the PESQ algorithm [21].

2.2 Virtual Speech Quality Objective Listener (ViSQOL)

ViSQOL is another intrusive objective speech quality metric that was developed by a team of researchers at Trinity College Dublin with the support of Google in 2015 to address limitations in existing metrics like PESQ, particularly for modern communication systems such as Voice over IP (VoIP) [7]. ViSQOL uses a model inspired by human auditory perception, making it suitable for both speech and general audio in narrowband and wideband signals.

It works by converting both the reference and degraded signals into spectrograms, which represent the signals across time and frequency. These spectrograms are divided into small patches, and a similarity score is computed for each patch using the Neurogram Similarity Index Measure (NSIM). To handle timing issues like clock drift or jitter, ViSQOL aligns the patches and compares warped versions of the reference patches [7]. The similarity scores are averaged and mapped to a MOS-LQO scale ranging from 1 to 5.

This mapping function relies on subjective datasets that were not publicly released [7]. Despite this limitation, ViSQOL has shown strong performance in predicting speech quality across various degradation types [23].

3 Methodology

This section details the methodology used to assess the quality of speech signals exposed to various degradation conditions across different languages. In the experiment, reference speech samples in multiple languages were subjected to various types of noise and progressively higher levels of degradation to compare the performance of objective speech quality metrics. This section is structured around the following parts: dataset pre-processing, degraded signal generation, degradation types, and used libraries and implementations.

3.1 Dataset Pre-Processing

The experiments were conducted using the ALLSSTAR Corpus Multilingual Dataset [24], which includes recordings of male and female speakers reading text samples in multiple languages. This dataset was selected because it contains samples in Turkish and Korean, two languages that were not previously used to evaluate the mapping functions to MOS-LQO scores [11]. Furthermore, it is open-source, promoting open science and making the experiment easy to reproduce, while adhering to ethical guidelines, which are further detailed in 5. The samples were recorded under the same conditions, ensuring data reliability and consistency for the research. The dataset also includes American English samples, which served as a baseline for comparison, given that English was used during the metric validation phase.

The dataset contains samples of speakers reading various pieces of text, such as the Declaration of Human Rights or individual sentences. For each language, 16 samples were extracted using the open-source Python library pydub [25], applying the *split_on_silence* method. The samples used have the following characteristics:

- The dataset consists of an equal male-female ratio (8 male, 8 female), with participants aged 18 to 29 years, and an average age of 22 years. All participants are native speakers of their respective languages. This balanced representation ensures that potential gender or age-related biases are minimized. However, some limitations were acknowledged and are discussed in 5.4.
- Each sample has a duration of 5-10 seconds, containing 1-2 sentences of recorded speech without any pauses in between, and an average speech activity rate of 97.8%. This ensures that the samples are long enough to capture meaningful speech characteristics with minimal interruptions, allowing for an analysis of the speech quality without irrelevant noise or silence.
- The audio samples are recorded in PCM signed 16-bit little-endian format with a bit rate of 353 kbps and a sampling rate of 22050 Hz. This high-quality audio format ensures that any observed degradation in speech quality can be accurately attributed to the added degradations rather than distortions from the recording itself.

3.2 Degradation Conditions

Various degradations were applied to simulate different real-world conditions for speech signals. These types of noise are commonly used in audio testing to replicate environments where speech quality may be affected by background interference [26]. The following list contains the degradation types used in the study:

- **Pink Noise:** A type of noise where the power density is inversely proportional to the signal frequency, resulting in a balanced sound across octaves [27]. Pink noise was used as it is commonly observed in nature, resembling sounds such as waterfalls, wind, and rain.
- **Blue Noise:** The opposite of pink noise, with power density increasing with frequency [27]. Blue noise was used to simulate distortion in systems that are sensitive to higher frequencies.
- **Babble Noise:** A type of noise that replicates the sound of human speech in a crowded environment, consisting of multiple overlapping voices. This is commonly encountered in settings such as busy restaurants, where conversations occur simultaneously, making it challenging to distinguish individual speech signals [28].

The noise samples used for speech degradation can be found on the project's GitHub repository [29] .

3.3 Degraded Signal Generation

To construct the degraded signal for evaluation by the full-reference metrics, a specific degradation process was applied to all original speech samples. The complete process is outlined in Figure 2.

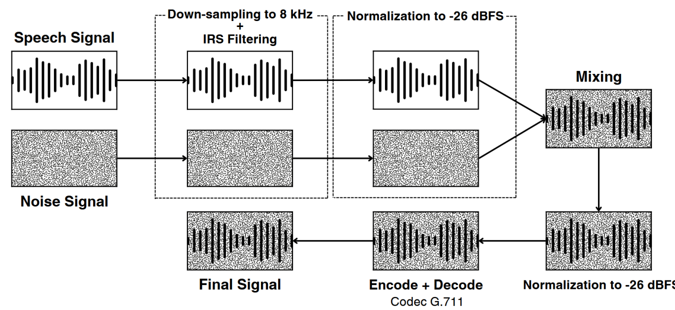

Figure 2: Diagram of the degradation process applied to the original speech samples for intrusive metric evaluation.

1. **Down-sampling and IRS filtering:** This step replicates the bandwidth limitations commonly found in legacy telecommunication systems. Initially, the speech and noise samples were converted to mono signals using the pydub function *split_to_mono*. They were then down-sampled to 8 kHz to achieve a wideband signal. Next, an Intermediate Reference System (IRS) filter [30] was applied, incorporating both low-pass and high-pass filtering. The IRS filter replicates the frequency response characteristics of older telecommunication systems by restricting the bandwidth to the range of 300 Hz to 3.4 kHz, resulting in a narrowband signal.
2. **Normalization:** Speech and noise signals were normalized to -26 dBFS (decibels relative to full scale) prior to mixing. This ensures consistent levels across all samples prior to the mixing stage and provides sufficient headroom to prevent clipping. The speech signal obtained during this step is used as the reference signal in the experiment.
3. **Mixing:** In the following stage, speech samples were mixed with the noise signals to introduce degradation effects. Noise samples were trimmed to match the length of the speech signals, and the noise intensity was varied using signal-to-noise ratio (SNR) levels ranging from -25 dB to 40 dB in increments of 5 dB. SNR measures the relative strength of the desired signal (speech) compared to the background noise, with higher SNR values indicating clearer speech and lower noise, while lower values reflect higher noise levels relative to the speech signal. The formula for calculating the required gain

to achieve the correct SNR for the noise signal is presented in equations 1 and 2, where RMS is the Root Mean Square, which is a measure of a signal's effective power or amplitude [31].

$$\mathrm{RMS}_{\mathrm{noise\_desired}} = \frac{\mathrm{RMS}_{\mathrm{signal}}}{10^{\frac{\mathrm{SNR}_{\mathrm{target}}}{20}}} \tag{1}$$

$$\mathrm{Gain}_{\mathrm{required}} = 20 \log_{10} \left(\frac{\mathrm{RMS}_{\mathrm{noise\_desired}}}{\mathrm{RMS}_{\mathrm{noise\_current}}} \right) \tag{2}$$

4. **Re-normalization:** The combined signal (speech + noise) was re-normalized to -26 dBFS to ensure consistent loudness levels across all samples after the mixing stage.

5. **Encoding/Decoding:** Samples were encoded and decoded using the ITU-T standard G.711 codec [32], specifically employing the A-law algorithm [33], which is generally used in telecommunication systems in Europe to reduce the dynamic range of the signal. This step simulates the effects of compression and decompression on speech quality, and results in the final degraded sample used for evaluation.

The final dataset consisted of 2016 degraded signals, derived from the following calculation: 14 SNR levels $\cdot$ 3 degradation types $\cdot$ 16 samples per language $\cdot$ 3 languages. These degraded signals were compared against a total of 48 reference signals (16 samples per language · 3 languages).

3.4 Used Libraries and Implementations

For the implementation, various open-source resources were used to assess and process speech signal quality. These resources were selected to support open science and ensure that the experiment can be freely reproduced by others. This is further detailed in 5.1.

- For PESQ, an open-source Python implementation from Wang et al. was used to evaluate narrowband speech at 8 kHz [34].
- The ViSQOL implementation provided by the Audio Toolbox in MATLAB 2024b was used in speech mode [35]. Due to the default input constraints of this specific implementation, it was required to resample the reference and degraded signals to 16 kHz prior to running the algorithm in the MATLAB engine for Python.
- Pydub [25] is a Python library designed for audio manipulation, which provides an interface to work with various audio formats. It was used during the degraded audio generation process.
- Numpy [36], skicit-learn [37], Scipy [38] and Pandas [39] are open-source Python libraries that were used to manipulate the results to perform statistical analyses and extract conclusions.
- Matplotlib [40] is a popular open-source Python library that was used to generate visualizations of the results.

4 Results

This section presents and discusses the results obtained from conducting the previously described experiment of evaluating PESQ and ViSQOL against Turkish, Korean and English degraded samples.

Figure 3 shows the evolution of the average PESQ and ViSQOL MOS-mapped scores per SNR value, segmented by language. These are the final scores generated by the algorithms after the final mapping function is applied, based on their original validation data. Note that only the final MOS-mapped score was returned from the algorithm, as the original scores were not available with the metric implementations. The data presented in this figure indicates that, while no significant difference is observed between the scores for PESQ, there is a notable trend in the ViSQOL results. Specifically, on average, ViSQOL mean scores for Turkish are 4.95% higher than those for English, corresponding to a difference of 0.19 points. However, when focusing on the SNR range from 0 dB to 25 dB, this difference increases to an average of 10.18%, equivalent to 0.34 points.

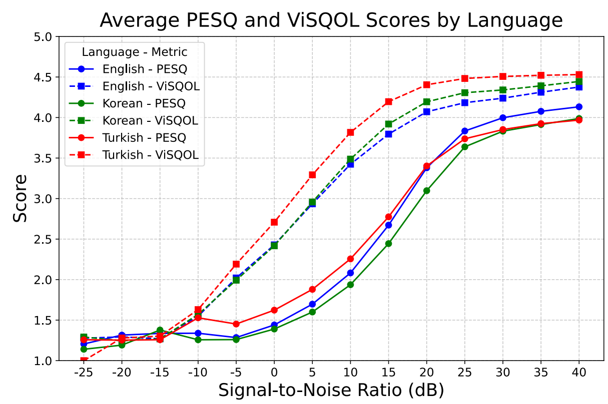

Figure 3: Evolution of average PESQ and ViSQOL MOS-mapped scores (1 = 'bad', 5 = 'excellent') across all degradation types, segmented by language, with SNR values from -25 dB to 40 dB.

This phenomenon is further illustrated in Figure 4, which presents a boxplot highlighting the key statistical features of the score distributions (median, mean, quartiles, minimum, maximum) for both PESQ and ViSQOL scores across languages. Notably, the median ViSQOL score for Turkish is observed to be 9% higher than those for English and Korean.

However, this information alone is insufficient to determine whether the differences in Turkish scores are statistically significant when compared to the rest. To further substantiate the findings and evaluate the influence of language on objective speech quality ratings, we use the Kolmogorov-Smirnov (KS) test [41], a non-parametric statistical method designed to assess whether two independent samples (PESQ and ViSQOL scores for different languages) are drawn from the same underlying distribution. This non-parametric test was chosen due to the bimodal nature of the distributions, as seen in Figure 5. Traditional parametric tests such as the t-test, Pearson correlation coefficient, or ANOVA are not appropriate, since they assume normality and unimodal distri-

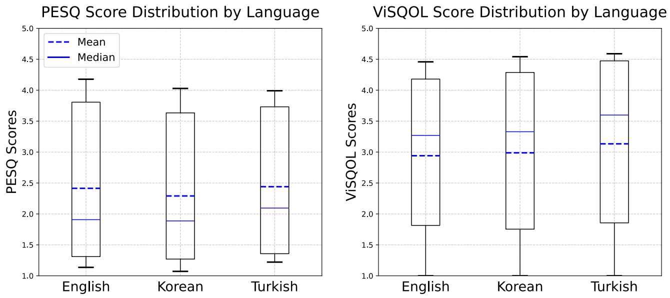

Figure 4: Boxplots of PESQ and ViSQOL scores (1 = 'bad', 5 = 'excellent') by language, showing the distribution of speech quality metrics.

butions [42]. In contrast, the KS test is suitable in this case because it is non-parametric, meaning it does not assume a specific data distribution and can be applied to any type.

The KS test compares the cumulative distribution functions (CDF) of two groups and calculates the maximum difference between them, resulting in a KS-statistic value. The corresponding p-value indicates the probability of obtaining a KS-statistic as large as the one observed, assuming that the null hypothesis is true. The null hypothesis for the KS test states that the two samples being compared come from the same distribution. If the p-value falls below a predefined significance threshold (α), we reject the null hypothesis and conclude that the distributions of scores differ significantly. A lower p-value means there is stronger evidence that the null hypothesis is incorrect.

For this study, a significance threshold of 0.05 was chosen, as it is commonly used to balance the risk of Type I errors (false positives), where a true null hypothesis is incorrectly rejected, while still ensuring practical significance [43] [44]. In contrast, for studies where the potential impact of false positives is larger, such as causing harm, a lower threshold is often applied to reduce this risk [45].

Table 1 presents the results of the KS tests for PESQ and ViSQOL scores. For the PESQ scores, the p-values for all pairwise comparisons are greater than 0.05. This means that we fail to reject the null hypothesis in all cases, indicating language does not appear to have a significant effect on the PESQ scores. In contrast, for the ViSQOL scores, a significant difference is observed in the comparison between English and Turkish, with a p-value of 0.02, which is below the 0.05 significance threshold. This suggests that the distributions of ViSQOL scores for English and Turkish differ significantly. However, the other pairwise comparisons both have p-values greater than 0.05. Therefore, based on the data analyzed, while there is a significant difference between the English and Turkish distributions of ViSQOL scores, no significant differences can be found in the other pairwise comparisons. It is important to note that failing to reject the null hypothesis does not prove that no differences exist, it only suggests that any observed differences could be due to random variation, and further investigations may be necessary.

Referring back to Figure 4, in the PESQ boxplot, the median lies below the mean for all languages, indicating that

Table 1: Kolmogorov-Smirnov (KS) test results for PESQ and ViSQOL scores segmented by language.

Metric	Comparison	KS-statistic	p-value
PESQ	English vs Korean	0.17	0.61
	English vs Turkish	0.20	0.44
	Korean vs Turkish	0.21	0.29
ViSQOL	English vs Korean	0.14	0.79
	English vs Turkish	0.33	**0.02**
	Korean vs Turkish	0.23	0.18

the values are clustered towards the lower end of the scale. In contrast, the ViSQOL boxplot shows the opposite case, where the median is positioned above the mean, suggesting that the values are clustered towards the higher end of the scale. Overall, the PESQ scores show a concentration of lower values, while the ViSQOL scores indicate a concentration of higher values.

++This is further illustrated by Figure 5, which displays a violin plot with the density distribution of PESQ and ViSQOL scores by language. For PESQ scores, the highest density is, on average, observed to have a score of 1.44, whereas, for ViSQOL scores, the peak density occurs on average at 4.19. As noted earlier, PESQ scores are more concentrated towards the lower end of the scale, resulting in a longer tail on the right. On the other hand, ViSQOL scores are more densely distributed at the higher end, leading to a longer tail on the left.

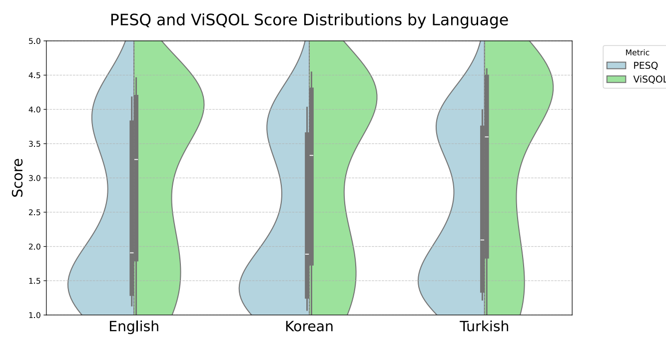

Figure 5: Density distributions of PESQ and ViSQOL scores (1 = 'bad', 5 = 'excellent') by language, illustrating the variation in score distributions across languages.

Figure 6 presents a comparison of the average PESQ and ViSQOL scores across different degradation types. Building on the previous discussion regarding the density distributions, this visualization further confirms that ViSQOL scores are generally higher than PESQ scores. This discrepancy between PESQ and ViSQOL scores is most noticeable in samples affected by babble noise. Specifically, the average difference in scores for babble noise is 72% greater than the average differences observed in pink and blue noise conditions. This suggests that background crowd noises have a smaller impact on ViSQOL scores for the considered languages.

To confirm this, Table 2 presents the KS test results for PESQ and ViSQOL scores across different degradation types. It can be observed that babble noise comparisons in ViSQOL

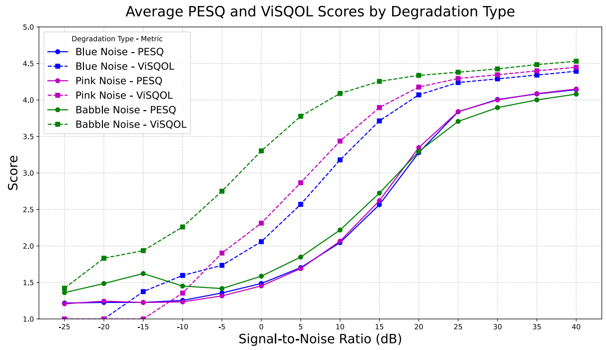

Figure 6: Evolution of PESQ and ViSQOL scores (1 = 'bad', 5 = 'excellent') segmented by degradation type with SNR values from -25 dB to 40 dB.

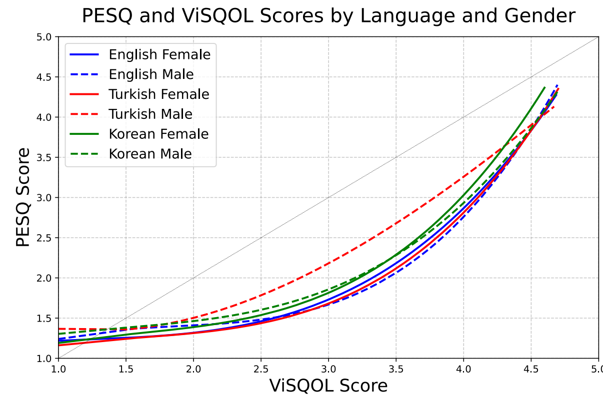

Figure 7: PESQ and ViSQOL scores (1 = 'bad', 5 = 'excellent') correlation segmented by gender, with data fitted using a cubic polynomial function.

produce the lowest p-values, with a significant difference between babble and blue noise scores (p-value = 0.04), assuming the same significance threshold as before of 0.05. Although the p-value for the pink and babble noise comparison is 0.06, which is close to significance, further investigation with additional speech samples and varied noise types is recommended for future work.

Table 2: Kolmogorov-Smirnov (KS) test results for PESQ and ViSQOL scores segmented by degradation type.

Metric	Comparison	KS-statistic	p-value
PESQ	Blue vs Pink Noise	0.12	0.93
	Blue vs Babble Noise	0.17	0.61
	Pink vs Babble Noise	0.19	0.44
ViSQOL	Blue vs Pink Noise	0.10	0.99
	Blue vs Babble Noise	0.31	**0.04**
	Pink vs Babble Noise	0.29	0.06

To analyze the impact of gender, Figure 7 illustrates the correlation between PESQ and ViSQOL scores, segmented by language and gender. The curves in the figure were generated by fitting a cubic polynomial function to the respective scores. As observed, all language groups exhibit a similar trend, with the exception of Turkish male speakers. The gap between PESQ and ViSQOL scores appears to be smaller for this group, indicating a distinct pattern.

To provide a more detailed analysis, Table 3 presents a statistical comparison of PESQ and ViSQOL scores across different languages and genders. The table includes key metrics such as the mean absolute deviation (MAD), root mean squared deviation (RMSD), and mean difference between PESQ and ViSQOL scores. Results are reported for three groups: the overall dataset, all subsets except Turkish male speakers (Non-TM), and Turkish male speakers (TM) specifically. Additionally, the last column shows the difference between the results of Turkish male speakers and the non-Turkish male speaker group, which is useful to highlight any variations specific to Turkish male speakers.

As seen, Turkish male speakers have lower MAD and RMSD values compared to both the overall dataset and the non-Turkish male group. Specifically, the MAD for Turkish male speakers is 0.62, which is 17.25% lower than the non-Turkish male group. Similarly, the RMSD for Turkish males is 0.77, 17.38% lower compared to the non-Turkish male group. Lastly, the mean difference between PESQ and ViSQOL scores is 0.18 lower for Turkish male speakers, representing a 37.9% increase in alignment compared to non-Turkish males.

Table 3: Statistical metrics: mean average deviation (MAD), root mean squared deviation (RMSD), and mean difference, comparing the correlation of PESQ and ViSQOL scores for the overall dataset, average scores excluding non-turkish male speakers (Non-TM), turkish male speakers (TM), and the difference between the last two groups (Diff).

Metric	Overall	Non-TM	TM	Diff
MAD	0.71	0.73	0.62	-0.11
RMSD	0.89	0.91	0.77	-0.13
Mean difference	-0.62	-0.65	-0.47	0.18

Since the mean difference is negative for both groups, PESQ scores are consistently lower than ViSQOL scores, but the gap is notably smaller for Turkish male speakers. This suggests that, on average, the PESQ and ViSQOL scores for Turkish male speakers are more closely aligned, whereas the non-Turkish male group experiences a larger discrepancy between the two metrics.

5 Responsible Research

This section outlines the key principles of responsible research that have guided this study. It emphasizes the importance of reproducibility, ethical conduct, and limitations throughout the research process.

5.1 Reproducibility of Research

This study follows the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [46] to ensure reproducibility and transparency. In alignment with these principles, all project code and data have been made publicly available on GitHub [29] under an MIT license [47]. Additionally, this paper provides a comprehensive explanation of the research methodology, including the degraded dataset generation steps, and detailed experimental results and figures.

All libraries and datasets used, such as PESQ, Pydub, Matplotlib, skicit-learn, Scipy, Pandas, and the ALLSSTAR Corpus dataset, are open-source, freely available to the public, and clearly referenced to facilitate reuse. Open-source resources were prioritized wherever possible. The only exception was the ViSQOL implementation used, which is accessible in Python only via the MATLAB engine. Although MATLAB is a paid platform, it is widely used in academia and is often accessible through educational institutions. For this reason, its use was considered appropriate for this research. By providing detailed documentation and openly sharing resources, this research enables independent validation and encourages scientific collaboration.

5.2 Ethical Considerations

The use of publicly available datasets like the ALLSSTAR Corpus ensures transparency and anonymity. All participants in the dataset were informed and gave consent for their contributions, with the dataset providing only demographic information like age, gender, and native language without any personal identifiers such as name. The dataset is licensed under a Creative Commons Attribution 4.0 International License [48], making it suitable for research purposes.

Efforts were made to avoid bias by ensuring an equal ratio of male to female participants and conducting a fair evaluation of speech quality across multiple languages. Furthermore, there were no financial or personal interests influencing this work, ensuring objectivity throughout the research.

5.3 Usage of Large Language Models (LLMs)

LLMs were used to simplify the research process by extracting information from text files, organizing it into tables in LaTeX format, and converting the data into JSON format. This made it easier to create visualizations, such as plots, which helped explain the results. The detailed prompts can be found in Appendix A.

5.4 Research Limitations

While the methodology was designed to provide robust and reliable results, it is important to acknowledge some limitations that may influence the generalizability of this study:

- The ALLSSTAR Corpus dataset includes a limited number of speakers per language, with most languages having fewer than eight female speakers. As a result, achieving a 50-50 male-to-female ratio was not feasible in all cases. To address this, additional samples from the same female speakers were used when necessary to maintain the desired gender balance.
- The dataset included both male and female voices, but its limited linguistic diversity (only two languages) and age range (18-29 years) may introduce gender-related and age-related biases, limiting generalizability to other groups.
- The study assumes uniform quality across all languages. Though Figure 3 shows that all languages converge at similar values, labeled subjective scores would be more optimal to confirm this assumption.
- While the degradation types and processes used were representative, they do not cover all possible real-world scenarios.
- An open-source version of PESQ was used, as it was the only available form for testing in Python in this study [34].
- Due to the lack of available datasets with subjective MOS quality ratings for the analyzed languages, including English, and recorded under the same conditions, only statistical analysis was possible to test the research question. However, the most reliable way to test the research hypothesis is to have such a dataset available. This is further detailed in Section 6.

6 Conclusions and Future Recommendations

To conclude, this study assessed the performance of PESQ and ViSQOL for predicting speech quality in Turkish and Korean, two languages outside their mapping function validation sets. While the results for both languages largely aligned with those for English, certain differences were observed. Specifically, Turkish exhibited larger ViSQOL scores, with an average 5% higher than English and Korean scores, and this difference increased to 10% in mid-range SNR values. Performing statistical tests, such as Kolmogorov-Smirnov, resulted in all the p-values exceeding the 0.05 significance threshold that was set for the experiment, except the one comparing ViSQOL scores for English and Turkish samples (0.02), indicating that there is a significant difference between the English and Turkish ViSQOL results.

Furthermore, it was found that ViSQOL scores tend to be higher and more concentrated at the upper end of the scale, while PESQ scores are more skewed towards the lower end. The impact of background noise varies, with babble noise showing the largest discrepancy between the two metrics for all languages, suggesting that ViSQOL is less sensitive to degradations involving background human speech. This was further confirmed by performing KS tests on the distributions, which revealed a significant difference between the ViSQOL scores for blue and babble noise, and a value of 0.06 for pink and babble noise, which is very close to the significance threshold and would be worth investigating in the future.

Finally, when analyzing gender-related effects, it was found that the Turkish-male speakers showcased the smallest gap between PESQ and ViSQOL scores, indicating a more consistent evaluation of speech quality within this subset, 37.9% more aligned compared to the other speaker groups.

Future research should aim to address the limitations discussed in subsection 5.4, as well as explore additional metrics such as POLQA. Expanding the study to include a wider variety of languages, age groups, and degradation types, as well as wideband audio would improve the relevance of the findings to real-world applications. It would be valuable to repeat the experiment using another language from the PESQ and ViSQOL validation set as a baseline for comparison, in order to strengthen the results.

The generalizability of objective speech quality metrics across languages relies heavily on having enough labeled data with subjective quality scores. This would help confirm biases toward certain languages and allow for adjustments and validation of the mapping functions based on each language's needs. Building such a dataset would be a step in the right direction for advancing future research in this area.

References

[1] M. Guéguin, R. L. Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," *EURASIP J. Adv. Signal Process.*, vol. 2008, 2008. [Online]. Available: https://doi.org/10.1155/2008/185248

[2] F. B. Ali, S. D. Larbi, M. Jaïdane, and K. Ridane, "Experimental mappings and validation of the dependence on the language of objective speech quality scores in actual GSM network conditions," in *17th European Signal Processing Conference, EUSIPCO 2009, Glasgow, Scotland, UK, August 24-28, 2009*. IEEE, 2009, pp. 2534–2538. [Online]. Available: https://ieeexplore.ieee.org/document/7077790/

[3] P. C. Loizou, "Speech quality assessment," in *Speech Enhancement: Theory and Practice*. CRC Press, 2011, pp. 623–654. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-19551-8_23

[4] International Telecommunication Union, "Methods for subjective determination of transmission quality," International Telecommunication Union (ITU), Recommendation P.800, August 1996, revised by ITU-T Study Group 12 (1993-1996). [Online]. Available: https://www.itu.int/rec/T-REC-P.800-199608-I/en

[5] ——, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001. [Online]. Available: https://www.itu.int/rec/T-REC-P.862

[6] ——, "Perceptual objective listening quality assessment (polqa)," Jan. 2011. [Online]. Available: https://www.itu.int/rec/T-REC-P.863

[7] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Visqol: The virtual speech quality objective listener," in *IWAENC 2012 - International Workshop on Acoustic Signal Enhancement, Proceedings, RWTH Aachen University, Germany, September 4th - 6th, 2012*. VDE-Verlag, 2012. [Online]. Available: http://www.vde-verlag.de/proceedings-de/453451035.html

[8] 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 *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 2127–2131. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-299

[9] HEAD Acoustics, "3quest: Applications & use in practice," HEAD Acoustics GmbH, Tech. Rep., Oct. 2010, application Note Rev0 (10/2008). [Online]. Available: https://cdn.head-acoustics.com/fileadmin/data/global/Application-Notes/Telecom/3QUEST-Applications-and-Use-In-Practice-Application-Note.pdf

[10] A. W. Rix, "Comparison between subjective listening quality and p.862 pesq score," Psytechnics Limited, Tech. Rep., 2003. [Online]. Available: https://www.sageinst.com/assets/download_doc/Subjective-Listening-Quality-And-P862-PESQ-Score-Comparison.pdf

[11] ITU-T, "Mapping function for transforming p.862 raw result scores to mos-lqo," International Telecommunication Union (ITU), Tech. Rep., Nov. 2003, iTU-T Recommendation P.862.1. [Online]. Available: https://www.itu.int/rec/T-REC-P.862.1/en

[12] U. of Pennsylvania, "About turkish language studies," Penn Language Center, 2024. [Online]. Available: https://web.sas.upenn.edu/turkish-studies/about/

[13] W3Techs, "Usage statistics of content languages for websites," 2025, accessed: January 23, 2025. [Online]. Available: https://w3techs.com/technologies/overview/content_language

[14] DataReportal, "Digital 2024: Turkey — datareportal – global digital insights," 2024, accessed on Thursday, January 23, 2025. [Online]. Available: https://datareportal.com/reports/digital-2024-turkey

[15] East Asian Languages and Civilizations, Harvard University, "Korean language overview," Harvard University Website, 2023. [Online]. Available: https://ealc.fas.harvard.edu/korean

[16] DataReportal, "Digital 2024: South korea — datareportal – global digital insights," 2024, accessed on Thursday, January 23, 2025. [Online]. Available: https://datareportal.com/reports/digital-2024-south-korea

[17] D. Konane, S. Tiemounou, and W. Y. S. B. Ouedraogo, "Impact of languages and accent on perceived speech quality predicted by perceptual evaluation of speech quality (pesq) and perceptual objective listening quality assessment (polqa): Case of moore, dioula, french and english," *Open Journal of Applied Sciences*, vol. 11, no. 12, pp. 1324–1332, 2021. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=114280

[18] H. Gamper, C. K. A. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2019, New Paltz, NY, USA, October 20-23, 2019*. IEEE, 2019, pp. 85–89. [Online]. Available: https://doi.org/10.1109/WASPAA.2019.8937202

[19] P. Manocha, Z. Jin, and A. Finkelstein, "Audio similarity is unreliable as a proxy for audio quality," in *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 3553–3557. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-405

[20] L. Malfait, J. Berger, and M. Kastner, "P.563—the itu-t standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006. [Online]. Available: https://ieeexplore.ieee.org/document/1709882

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings*. IEEE, 2001, pp. 749–752. [Online]. Available: https://doi.org/10.1109/ICASSP.2001.941023

[22] Psytechnics Limited, "Pesq: An introduction," Psytechnics Limited, Ipswich, Suffolk, United Kingdom, Tech. Rep., 2004, white Paper. [Online]. Available: https://www.sageinst.com/assets/download_doc/1639138674-wp_pesq_introduction.pdf

[23] A. Hines, E. Gillen, and N. Harte, "Measuring and monitoring speech quality for voice over IP with polqa, visqol and p.563," in *16th Annual Conference of the International Speech Communication Association, INTERSPEECH 2015, Dresden, Germany, September 6-10, 2015*. ISCA, 2015, pp. 438–442. [Online]. Available: https://doi.org/10.21437/Interspeech.2015-171

[24] A. R. Bradlow, "Allsstar: Archive of l1 and l2 scripted and spontaneous transcripts and recordings," Northwestern University Speech Communication Research Group, n.d. [Online]. Available: https://speechbox.linguistics.northwestern.edu/allsstar

[25] J. Robert, "Pydub: Manipulate audio with a simple and easy high-level interface," 2018, version 0.25.1. Accessed: January 25, 2025. [Online]. Available: https://github.com/jiaaro/pydub

[26] H. C. Stronks, J. J. Briaire, and J. H. M. Frijns, "The temporal fine structure of background noise determines the benefit of bimodal hearing for recognizing speech," *JARO - Journal of the Association for Research in Otolaryngology*, vol. 21, no. 6, pp. 527–544, October 2020. [Online]. Available: https://doi.org/10.1007/s10162-020-00772-1

[27] D. Yan, J. Guo, B. Wang, X. Zhang, and P. Wonka, "A survey of blue-noise sampling and its applications," *J. Comput. Sci. Technol.*, vol. 30, no. 3, pp. 439–452, 2015. [Online]. Available: https://doi.org/10.1007/s11390-015-1535-0

[28] N. Krishnamurthy and J. H. L. Hansen, "Babble noise: Modeling, analysis, and applications," *IEEE Trans. Speech Audio Process.*, vol. 17, no. 7, pp. 1394–1407, 2009. [Online]. Available: https://ieeexplore.ieee.org/document/5175669

[29] J. P. Lopez, "Performance of objective speech quality metrics on languages beyond validation data," GitHub Repository, January 2025. [Online]. Available: https://github.com/javipeloza/bias-speech-quality-metrics

[30] International Telecommunication Union, "Specification for an intermediate reference system," ITU-T, Geneva, Switzerland, Recommendation P.48, Nov. 1989. [Online]. Available: https://www.itu.int/rec/T-REC-P.48/en

[31] Z. Yao, "Signal to noise ratio, kurtosis," *Journal of Analytical and Bioanalytical Techniques*, vol. 14, no. 2, p. 497, 2023, this is an open-access article distributed under the terms of the Creative Commons Attribution License. [Online]. Available: https://www.omicsonline.org/open-access-pdfs/signal-to-noise-ratio-kurtosis.pdf

[32] International Telecommunication Union, "Itu-t recommendation g.711.0: Lossless compression of g.711 pulse code modulation," International Telecommunication Union (ITU), Tech. Rep., September 2009. [Online]. Available: https://www.itu.int/rec/T-REC-G.711.0-200909-I/en

[33] Texas Instruments, "A-law and μ-law companding implementations using the tms320c54x," Texas Instruments, Tech. Rep., October 1997, application Report SPRA163A. [Online]. Available: https://www.ti.com/lit/an/spra163a/spra163a.pdf

[34] M. Wang, C. Boeddeker, R. G. Dantas, and A. Seelan, "Pesq (perceptual evaluation of speech quality) wrapper for python users," May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6549559

[35] MathWorks, Inc., "Visqol: Objective metric for perceived audio quality," MathWorks Documentation, January 2024. [Online]. Available: https://www.mathworks.com/help/audio/ref/visqol.html

[36] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, September 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and

É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html

[38] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, pp. 261–272, February 2020. [Online]. Available: https://doi.org/10.1038/s41592-019-0686-2

[39] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 51–56. [Online]. Available: https://doi.org/10.25080/Majora-92bf1922-00a

[40] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, May 2007. [Online]. Available: https://ieeexplore.ieee.org/document/1709882

[41] F. J. M. Jr., "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951. [Online]. Available: https://www.jstor.org/stable/2280095

[42] M. W. Fagerland, L. Sandvik, and P. Mowinckel, "Parametric versus non-parametric tests of location in biomedical research," *Statistics in Medicine*, vol. 31, no. 14, pp. 1350–1367, 2012. [Online]. Available: https://doi.org/10.1002/sim.4385

[43] R. A. Fisher, *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver and Boyd, 1925. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4612-4380-9_6

[44] B. M. Cesana, "What p-value must be used as the statistical significance threshold? p¡0.005, p¡0.01, p¡0.05 or no value at all?" *Biomedical Journal of Scientific & Technical Research*, vol. 6, no. 3, pp. 5310–5313, 2018. [Online]. Available: https://biomedres.us/pdfs/BJSTR.MS.ID.001359.pdf

[45] G. Di Leo and F. Sardanelli, "Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach," *European Radiology Experimental*, vol. 4, p. 18, 03 2020. [Online]. Available: https://www.researchgate.net/publication/339844197_Statistical_significance_p_value_005_threshold_and_applications_to_radiomics-reasons_for_a_conservative_approach

[46] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, and J. Zhao, "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, 2016. [Online]. Available: https://doi.org/10.1038/sdata.2016.18

[47] Massachusetts Institute of Technology, "The mit license," Open Source Initiative, 1988, accessed: January 20, 2025. [Online]. Available: https://opensource.org/license/mit/

[48] Creative Commons, "Attribution 4.0 international (cc by 4.0) license," Creative Commons License Deed, 2013, accessed: January 25, 2025. [Online]. Available: https://creativecommons.org/licenses/by/4.0/

A Use of Large Language Models (LLMs):

This section contains a list of the prompts that were used to create tables in this document:

Question: Convert the following information in this txt file [*pairwise_statistical_test_results.txt*] to a clear table in Latex. The PESQ and ViSQOL labels should span the three rows they have in common.

Answer: Here is the LaTeX code for the table based on your provided information:

```
\begin{table}[h!]
\centering
\begin{tabular}{|c|c|c|}
\hline
\multirow{3}{*}{PESQ} & Statistic & p-value \\ \cline{2-3}
                      & English vs Turkish & 0.19 & 0.44 \\ \cline{2-3}
                      & English vs Korean   & 0.17 & 0.61 \\ \cline{2-3}
                      & Turkish vs Korean   & 0.21 & 0.29 \\ \hline
\multirow{3}{*}{ViSQOL} & English vs Turkish & 0.33 & 0.02 \\ \cline{2-3}
                        & English vs Korean   & 0.14 & 0.79 \\ \cline{2-3}
                        & Turkish vs Korean   & 0.24 & 0.19 \\ \hline
\end{tabular}
\end{table}
```

Question: From the following txt files [*deviation_metrics.txt*, *turkish_male_vs_others_comparison_results.txt*], extract the following information to a clear table in Latex: 3 columns (overall, non turkish male, turkish male). For each column, extract the following: Mean average deviation, root mean squared deviation, bias (mean difference PESQ, ViSQOL).

Answer: Here is the LaTeX code to create a table based on the data you provided. The table will include three columns (Overall, Non-Turkish Male, Turkish Male) and display the requested metrics.

```
\begin{table}[h!]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
\textbf{Metric} & \textbf{Overall} & \textbf{Non-Turkish Male} & \textbf{Turkish Male} \\ \hline
Mean Average Deviation (MAE) & 0.7109 & 0.7288 & 0.6214 \\ \hline
Root Mean Squared Deviation (RMSE) & 0.8866 & 0.9063 & 0.7721 \\ \hline
Bias (Mean Difference PESQ - ViSQOL) & -0.6232 & -0.6531 & -0.4736 \\ \hline
\end{tabular}
\end{table}
```

Question: I have txt files with experiment results that should be JSON but have np.int64(number) and single quotes instead of double. Write a Python script to fix this and save the corrected JSON files in a json_fixed folder.

Answer: Here's a Python script that will fix the JSON formatting in .txt files and save the corrected JSON files:

```
def fix_experiment_results(file_path: str):
    with open(file_path, "r", encoding="utf-8") as f:
        content = f.read()

    # Replace np.int64(NUMBER) with just NUMBER
    content = re.sub(r"np\.int64\((\d+)\)", r"\1", content)

    # Replace single quotes with double quotes for valid JSON
    content = content.replace("'", '"')

    # Parse to ensure valid JSON
    try:
        data = json.loads(content)
    except json.JSONDecodeError as e:
        print("Error fixing JSON:", e)
        return

    fixed_file_path = file_path.replace(".txt", "_fixed.json")
    with open(fixed_file_path, "w", encoding="utf-8") as f:
        json.dump(data, f, indent=4)
```