Performance of Objective Speech Quality Metrics on Languages Beyond Validation Data

EEMCS, Delft University of Technology

Author: Javier Pérez López | j.perezlopez@student.tudelft.com

Supervisors: Jorge Martinez Castañeda, Dimme de Groot

1. Introduction

- The measurement of **speech quality** in telecommunication systems is essential for optimal user experience and has been extensively studied to adapt to evolving technologies. Over time, speech quality metrics have been developed and **standardized** by the ITU-T to enable consistent assessment [1].
- **Subjective metrics** rely on human listeners rating the perceived quality of speech signals on a standardized scale, resulting in a Mean Opinion Score (MOS) (1=“bad, 5=”excellent”). These methods are costly, **time-consuming**, and challenging to scale.
- **Objective metrics** offer an automated approach to predict speech quality, aiming to replicate human ratings. They are significantly **faster**, more scalable, and cost-effective.
- These metrics, developed using datasets from a **limited set of languages**, internally produce a score that is mapped to a MOS for consistent evaluation. This mapping process, relying on limited data, may affect performance for languages outside the initial validation set [2].
- The rapid expansion of **multilingual speech** technologies has raised concerns about the need to ensure speech quality metrics are robust across diverse languages for fair and accurate assessments.
- We benchmark **PESQ** and **ViSQOL** (two intrusive speech quality metrics) by comparing their performance on **Turkish** and **Korean** speech samples to that of **English**, across various degradation conditions.

Research Question:

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?

References:

- **[1]** 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
- **[2]** 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. https://www.itu.int/rec/T-REC-P.862.1/en

2. Methodology

Reference speech samples were degraded with varying noise types and severity levels to evaluate speech quality metrics.

Dataset:

- This study used the open ALLSSTAR Corpus Multilingual Dataset, featuring 16 samples per language, each consisting of 5-10 seconds of continuous speech from individual native speakers (8 male, 8 female), aged 18-29.

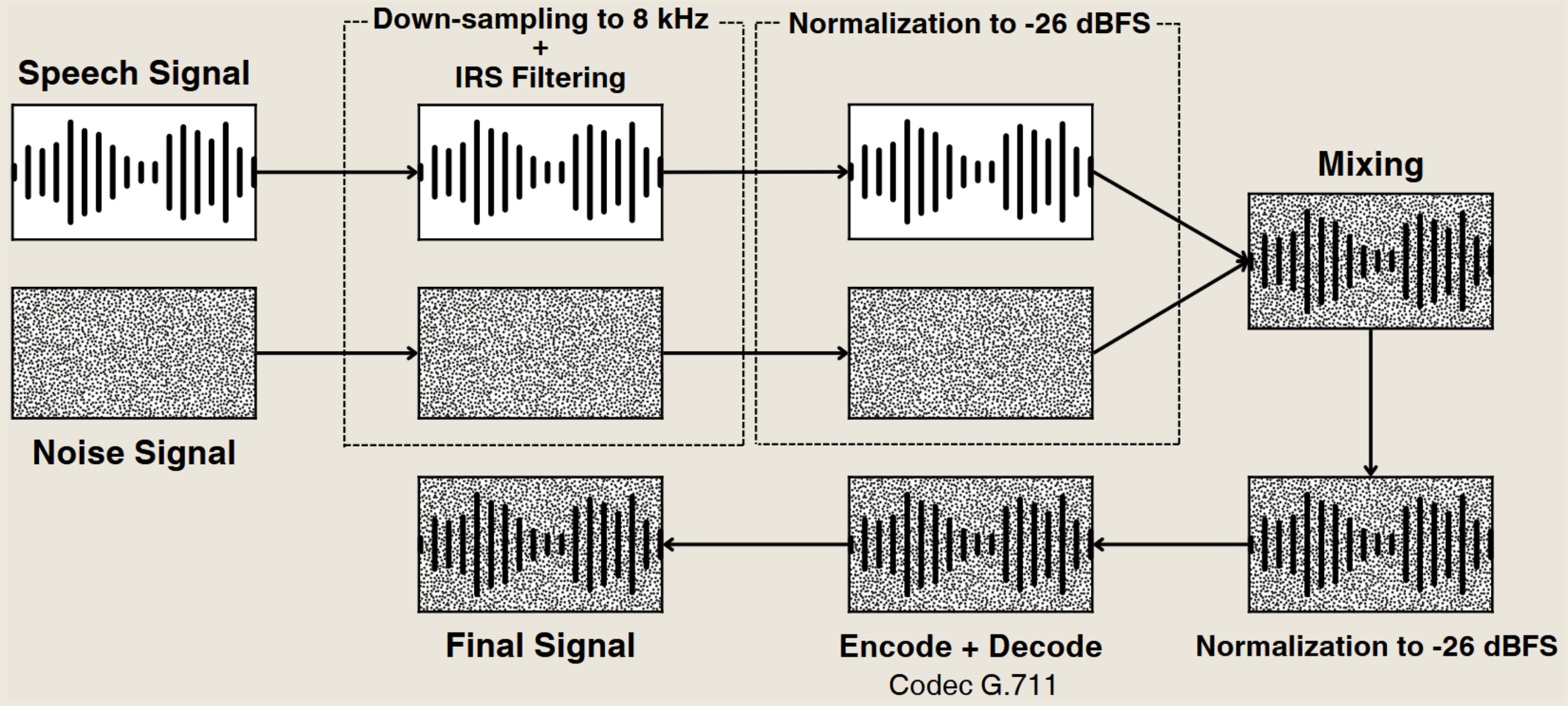
Degradation Conditions:

- **Pink Noise:** Noise with power density inversely proportional to frequency, resembling natural sounds like waterfalls and rain.
- **Blue Noise:** Noise with increasing power density at higher frequencies, used to simulate high frequecy distortions.
- **Babble Noise:** Overlapping human speech, common in crowded environments like restaurants.

Degraded Signal Generation:

All original samples were processed to simulate degradations in telecommunication systems:

1. **Down-sampling**: Down-sampling to 8 kHz (wideband) and IRS filtering to obtain a narrowband sample (300 Hz - 3.4 kHz).
2. **Normalization**: Speech and noise signals were normalized to -26 dBFS to ensure consistency, prevent clipping, and serve as the **reference signal** in the experiment.
3. **Mixing**: Speech samples were mixed with noise signals at Signal-to-Noise-Ratio (SNR) levels from -25 dB to 40 dB. Higher SNR values indicate clearer speech, while lower values reflect more noise.
4. **Normalization of Mixed Signal**: The combined signal (speech + noise) was normalized again to -26 dBFS.
5. **Encoding/Decoding**: The final signal was encoded and decoded using the G.711 codec with the A-law algorithm to simulate telecommunication compression effects for evaluation.



3. Results

- The results in Figure 1 indicate that **Turkish** scores were on **average 5% higher** than English, increasing to **10%** in mid-range SNR values. According to the Kolmogorov-Smirnov (KS) tests in Table 1, language did not have a significant impact on PESQ scores. However, **significant differences** were found between **Turkish** and **English** scores, and between **blue** noise and **babble noise**.

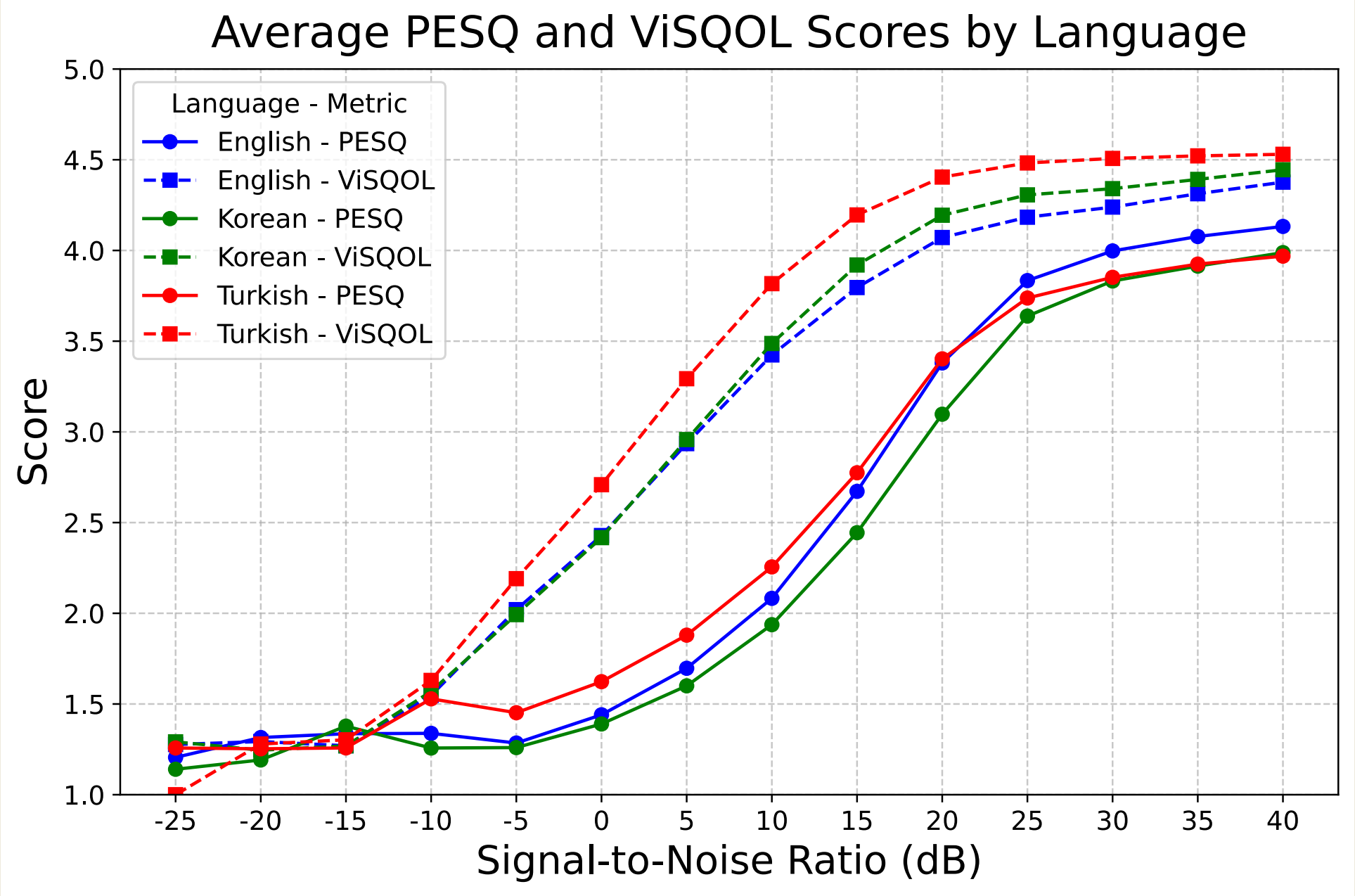


Figure 1: Average PESQ and ViSQOL scores by language

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

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

Table 1: Kolmogorov-Smirnov (KS) tests by language and noise

- PESQ distributions in Figure 2 show that scores are clustered towards the **lower end** (1.44), while ViSQOL shows the opposite, towards the **higher end** (4.19). Additionally, the median ViSQOL score for **Turkish** was **9% higher** than for English and Korean.

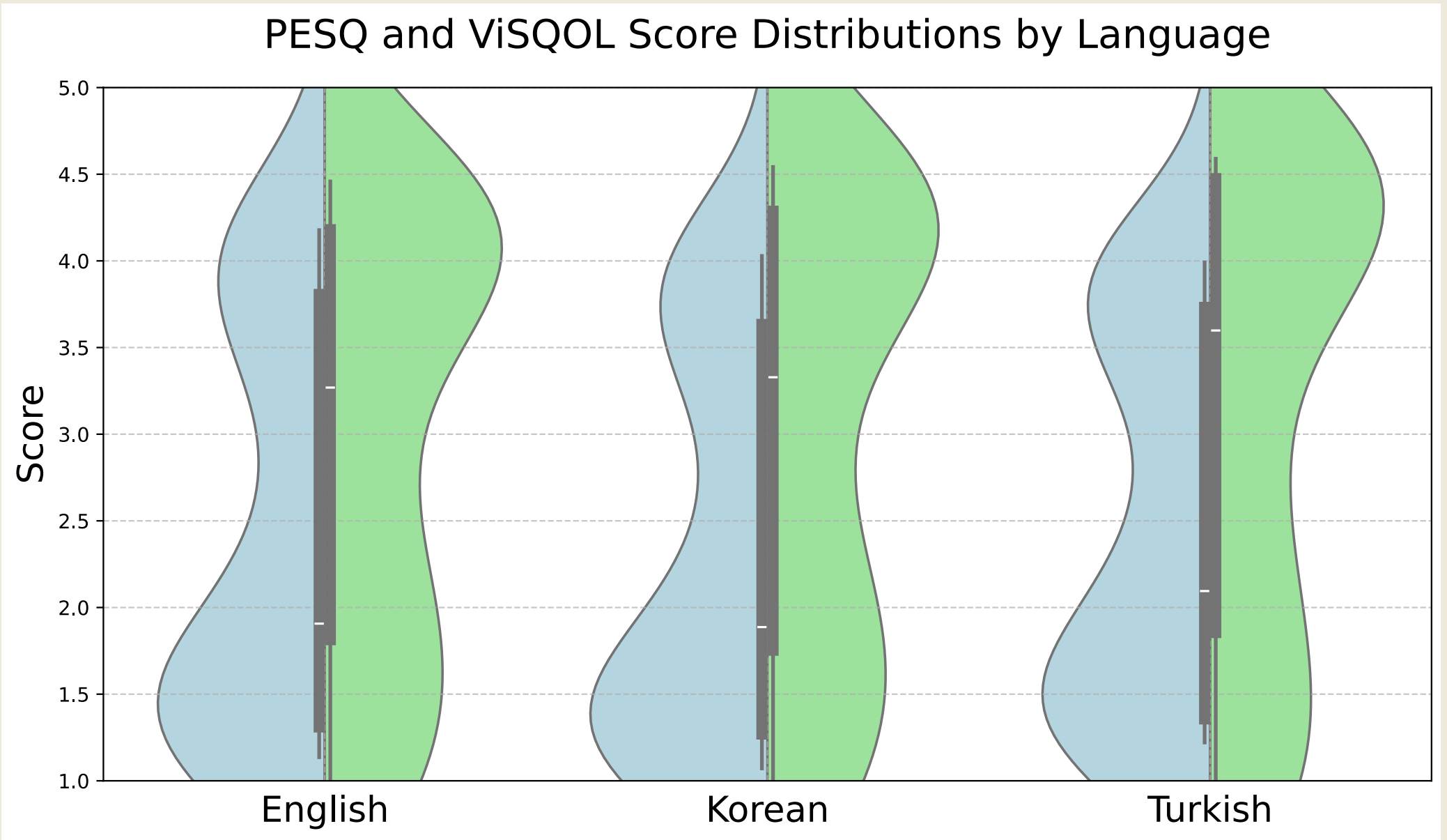


Figure 2: Score Distributions by Language (blue = PESQ, green = ViSQOL)

- Figure 3 shows that, except for Turkish male speakers, all language groups exhibited a similar trend in correlation between PESQ and ViSQOL scores. **Turkish male** speakers had a smaller gap between the two metrics, with **lower deviation** values compared to the rest of groups. The results indicated that PESQ and ViSQOL scores were **37.9%** more closely **aligned** for Turkish male speakers.

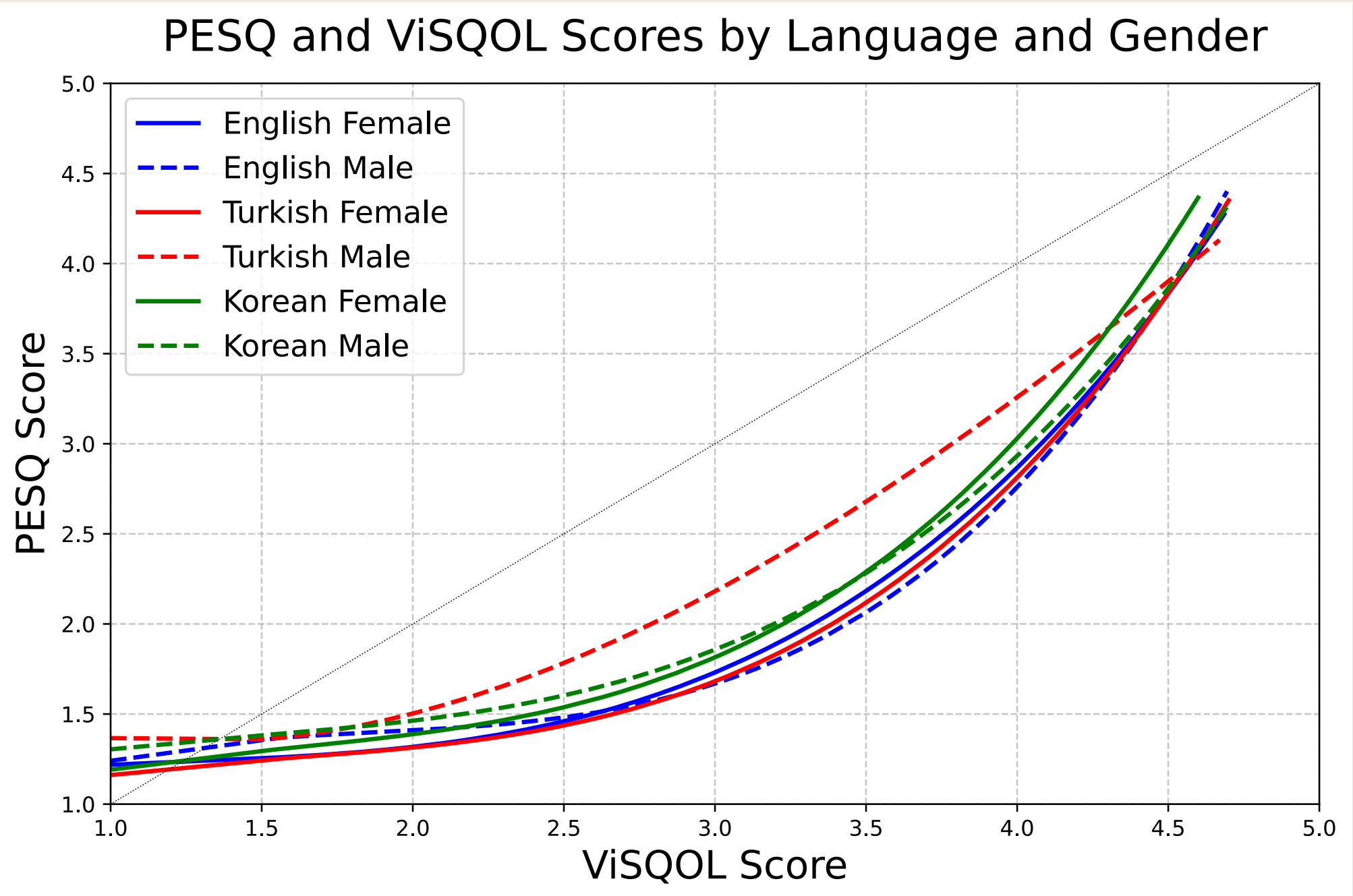


Figure 3: Metric correlation by language and gender

4. Conclusions & Future Work

- In conclusion, Turkish exhibited higher ViSQOL scores than English and Korean, with significant differences between Turkish and English ViSQOL scores, and Turkish male speakers showing the highest correlation between both metrics. It was also found that babble noise degradations had a lower impact on ViSQOL.
- As future work, it is recommended to expand the research with more languages and metrics, and create a new dataset with labeled subjective scores.