An Exploratory Examination of Objective Intelligibility Metrics Under Reverberant Conditions

Mingyi Jin[1]

Supervisor(s): Jorge Martinez Castaneda[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: Mingyi Jin
Final project course: CSE3000 Research Project
Thesis committee: Jorge Martinez Castaneda, Dimme de Groot, Przemyslaw Pawelczak

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Clear communication in public address systems is essential, especially in environments where safety or information clarity is critical. Speech intelligibility is often assessed using objective intelligibility metrics (OIMs), which predict intelligibility through mathematical models. These metrics perform well in non-highly reverberant conditions but face challenges in highly reverberant environments and with non-European languages like Mandarin. This study examines the performance of three intrusive OIMs-ESTOI, HASPI, and SIIB$_{\text{Gauss}}$-in two aspects: (1) how these metrics perform under different reverberation conditions for English, using STIPA as a reference, and (2) how robust these metrics are by comparing the variances of scores between Mandarin and English. The results show that the variances of predicted scores by the test metrics are equal between Mandarin and English. HASPI, ESTOI, and SIIB$_{\text{Gauss}}$ demonstrate similar performance across a broader range of reverberation conditions (from a T60 of 0.05s to 7s) for English, contradicting the theory that most intrusive intelligibility metrics struggle with severe reverberation conditions [1]. The findings highlight the need for further research to evaluate potential biases in OIMs and their performance across languages. Incorporating listening tests could provide a more solid examination of these metrics under diverse conditions for different languages.

1 Introduction

Public address (PA) systems are electronic systems that combine a mixer, amplifier, and speakers to deliver messages to a crowd. Effective communication in PA systems is critical, especially in environments where safety or clarity of information is paramount, such as transportation hubs, educational institutions, and emergency settings. Measuring speech intelligibility in these systems is an important task, ensuring that messages are comprehensible across diverse conditions.

Speech intelligibility is commonly defined as the percentage of words a listener can accurately recognize [2], and its assessment can be broadly categorized into subjective and objective methods. Subjective metrics rely on human listeners who evaluate the intelligibility of speech in controlled conditions, often in the form of word intelligibility tests and sentence intelligibility tests. These methods directly reflect human perception but are time-consuming and costly. Objective intelligibility metrics (OIMs), on the other hand, use mathematical models to predict intelligibility. They can be divided into intrusive and non-intrusive approaches. Intrusive methods, also known as reference-based approaches, rely on a clean speech or noise sample as a reference. These methods compare the degraded signal to the reference to calculate the intelligibility score [3]. Non-intrusive metrics analyze degraded signals without the need for reference signal. Research shows that intrusive metrics are more correlated with intelligibility than non-intrusive metrics[4], and non-intrusive metrics are more advantageous in situations where reference signal is not available or real-time monitoring is needed.

Early OIMs, such as the Articulation Index (AI) [5] and later the Speech Intelligibility Index (SII) [6], focused on analyzing the contribution of signal-to-noise ratios across frequency bands to predict intelligibility. These models laid the groundwork for the STI [7], which uses the modulation transfer function (MTF) to account for distortions in time and frequency caused by noise, reverberation, echo, and non-linear effects like peak clipping [8]. Measuring STI involves a modulated Gaussian noise test signal to assess how well modulation depth is preserved after passing through a transmission channel or acoustic space. The STI has been validated for English by comparing its scores with subjective speech intelligibility scores across a range of conditions, including reverberant environments, using various intelligibility assessment methods [9] [10]. However, STI takes 15 minutes [9] to calculate, making it time-consuming. STIPA, which only takes 15 seconds [9] measurement time, was later developed as a simpler and faster alternative, designed specifically for testing public address systems [11].

Despite the outstanding performance of STI and STIPA under reverberant conditions, this is not the case for other speech-input based intrusive intelligibility metrics. T60 (reverberation time) is defined as the time it takes for a sound to decay by 60 dB and reflects the severity of reverberant distortion. A study showed that while intrusive metrics such as ESTOI (the Extended Short-Time Objective Intelligibility Measure) [12] and SIIB (Speech Intelligibility in Bits) [13] performed well at low T60 values, all other metrics tested in that study except HASPI (Hearing-Aid Speech Perception Index) [14] demonstrated poor performance at a T60 of 1 second. The author proposed that this occurs because many intrusive intelligibility metrics rely on time alignment between clean and degraded signals, making them overly sensitive to temporal blurring caused by severe reverberant distortion [15].

Another notable aspect is that, despite their widespread use, challenges persist in applying OIMs to languages that were not considered during their development phase. These metrics were primarily developed using linguistic and phonetic properties specific to non-tonal European languages [5]. For example, an improved STI method proposed by Lin, which takes characteristics of Mandarin speech into account, shows better correlation with subjective intelligibility compared to original speech-input based STI method [16].

A study on the intelligibility of English, Polish, Arabic, and Mandarin using **subjective tests** showed that, under the same room acoustic conditions, English was the most intelligible language in both noisy and reverberant environments. The study also found that significant differences in subjective intelligibility can arise between languages, particularly in acoustically challenging spaces [1]. This raises the question of whether OIMs developed based on European languages can accurately assess the speech intelligibility of other languages. Mandarin, as a tonal language, relies on pitch and tonal variation for meaning. These linguistic differences could influence how intelligibility is perceived and interpreted, raising concerns about potential linguistic bias in metrics like ESTOI, $\text{SIIB}_{\text{Gauss}}$ and HASPI, which have not been tested for Mandarin under severely reverberant conditions.

This research is an exploratory examination of objective intelligibility metrics, focusing on the following main research question and subquestions:

Research Question: How do ESTOI, $\text{SIIB}_{\text{Gauss}}$ and HASPI perform under different reverberant conditions?

- Sbquestion 1: How do ESTOI, $\text{SIIB}_{\text{Gauss}}$ and HASPI perform under different reverberant conditions for English?
- Subquestion 2: How robust are ESTOI, $\text{SIIB}_{\text{Gauss}}$ and HASPI, for Mandarin compared to English under reverberant conditions?

As part of the methodology, we use STIPA, which has been thoroughly tested for English under reverberation [9], as a reference metric to evaluate how other test OIMs correlate with it under different reverberation conditions for English. To assess robustness across languages, variances of predicted scores at each T60 are compared to identify any differences between Mandarin and English.

2 Objective Intelligibility Metrics

As mentioned earlier, the study [15] suggested that the reason HASPI performed well at a T60 of 1 second, while other metrics such as ESTOI and $\text{SIIB}_{\text{Gauss}}$ only performed poorly, is due to the severe distortion caused by longer reverberation times. To investigate this claim, we selected ESTOI, $\text{SIIB}_{\text{Gauss}}$, and HASPI as test metrics for comparison with STIPA at a broader T60 range for English.

2.1 Reference Metric for English: STIPA

The traditional STI process [9] revolves around analyzing the degradation of modulated test signals designed to replicate the spectral and temporal characteristics of human speech. These test signals, passed through the communication channel or environment under evaluation, experience modifications due to the system's characteristics. By examining how the modulation in these signals is altered using Modulation Transfer Function. The measurement process considers seven octave frequency bands and evaluates modulation frequencies within these bands. By computing the effective signal-to-noise ratio (SNR) for each combination of octave band and modulation frequency, the STI aggregates the results through a weighted summation process. This ensures that the method reflects the cumulative impact of frequency-specific degradations and their contributions to speech intelligibility. STIPA [9]is a simplified version of STI tailored for public address systems. Instead of analyzing 14 modulation frequencies per octave band as in full STI, STIPA focuses on a smaller subset, reducing the measurement time to approximately 15 seconds without significantly compromising accuracy. This efficiency is particularly valuable when evaluating PA systems in real-world environments, such as transportation hubs, auditoriums, and open public spaces. The STIPA implementation from this GitHub page [17] is used, and it is distributed under GPL-3.0 License.

2.2 Test Metrics

The Extended Short-Time Objective Intelligibility (ESTOI)

ESTOI [12] is an enhancement of the Short-Time Objective Intelligibility (STOI) [18] algorithm, designed to predict speech intelligibility in environments with highly modulated noise sources or non-linear distortions. While STOI assumes that frequency bands contribute to intelligibility independently, ESTOI extends STOI by incorporating spectral correlations between frequency bands and accounting for correlations between the temporal envelopes of clean and noisy speech signals. ESTOI has been shown to perform well under reverberant conditions where T60 is less than 1 second [15]. The ESTOI implementation used is from this Github page [19], which has MIT License.

Speech Intelligibility in Bits (SIIB) Gaussian

SIIB [13] is an information-theoretic metric that estimates speech intelligibility by quantifying the amount of information shared between clean and distorted speech signals, measured in bits per second. It stands out for its conceptual simplicity, solid theoretical foundation, and strong performance. $\text{SIIB}_{\text{Gauss}}$ [15] simplifies the mutual information estimation process in SIIB by leveraging the information capacity of a Gaussian communication channel. It is computationally faster while maintaining performance levels comparable to SIIB. Similar to ESTOI, SIIB and $\text{SIIB}_{\text{Gauss}}$ have demonstrated strong performance in reverberant conditions with T60 values below 1 second. In the experiments, we use the $\text{SIIB}_{\text{Gauss}}$ implementation from the pySIIB library, which is distributed under the MIT License.

The Hearing-Aid Speech Perception Index (HASPI)

HASPI [20] predicts speech intelligibility for both normal-hearing and hearing-impaired individuals using an auditory model that accounts for hearing loss. It compares the envelope and temporal fine structure outputs of a reference signal to those of a test signal. In this study, HASPI was the only tested intrusive intelligibility metric that performed well under T60 of 1 second, which is considered to be rather severe reverberant distortion. Version 2 [14] of HASPI also demonstrated significantly better performance compared to version 1 for speech in reverberation. In our experiments, we use the HASPI version 2 implementation included in the Clarity Python library, which has MIT License [21].

3 Experimental Setup

In this section, we discuss the datasets used, the process of generating degraded signals, and the experimental procedures in our data-driven approach to address the research questions.

3.1 Materials

To ensure that the types of clean signals are similar to those played on PA systems, the clean signals should be meaningful monologues. Additionally, to allow for the meaning comparison of score variances at each T60, the clean signals should be of high recording quality. Based on these two criteria, we selected the following two datasets for Mandarin and English, respectively.

For the English dataset, the TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) is used. TIMIT contains recordings from 630 speakers (70 percent male and 30 percent female), each reading phonetically rich sentences of approximately three seconds. The speakers are from eight dialect regions in the USA. These sentences are designed to provide speech data for acoustic-phonetic studies and the development and evaluation of automatic speech recognition systems. Each speech utterance is a single-channel waveform file with a 16 kHz sampling rate and 16-bit depth.

For Mandarin, the AISHELL-3 dataset [22] is used, which includes recordings from 218 speakers (42 male and 176 female) reading smart home voice commands, news reports, geographic information, and number strings in Mandarin Chinese of various lengths. Regarding the speakers' accents, 165 have a northern accent, 51 have a southern accent, and 2 have other accents. Each speech utterance is a single-channel waveform file with a 44.1 kHz sampling rate and 16-bit depth.

Although the content of the sentences in TIMIT and AISHELL-3 may differ, they are all emotion-neutral, meaningful monologues of high recording quality, ensuring consistency between the two datasets. The difference in sampling rates is addressed by resampling the 44.1 kHz waveform files to a 16 kHz sampling rate.

From each of the TIMIT and AISHELL-3 datasets, 84 speakers (42 male and 42 female) were selected, and for each speaker, three utterances of three seconds in length were chosen. This resulted in one Mandarin subset and one English subset, each containing 84×3=252 clean utterances. For STIPA, a 15 second long test signal was generated using the and resampled to 16 kHz to match the sample rate of the datasets. Because AISHELL-3 only has 42 male speakers and we wanted to ensure a balance of sexes in our samples, we selected an equal number of male and female speakers. Despite not fully representing the entire population of Mandarin and English speakers, these signals should largely mitigate the effects of individual variation.

3.2 Generating Degraded Signals

To simulate real-world acoustic degradation in public spaces, each clean audio signal (1 STIPA test signal, 252 Mandarin sentences, and 252 English sentences) was convolved with 40 Room Impulse Responses (RIRs) of the same T60 value to generate reverberation-degraded signals. Each T60 value corresponds to one room type, with T60 values ranging from 0.05s, 0.17s, 0.31s, 0.48s, 0.71s, 1.17s, 1.92s, 3.15s, and 7.00s, resulting in a total of 40 x 9 = 360 different conditions. The T60 values are denser in the lower range because the human ear is more sensitive to changes in lower T60 ranges. This range was chosen to reflect T60 conditions similar to those described in [9], where STIPA was tested. These degraded signals are crucial for testing as they simulate a variety of real-world environments in which we intend to compare test metrics with STIPA. Public spaces vary significantly in their acoustic properties, and these degraded signals aim to reflect this diversity. The T60 values cover typical acoustic conditions found in various spaces, ranging from classrooms and offices (smaller T60 values) to larger, more reverberant spaces such as airports, stations, and tunnels (larger T60 values). The RIRs were generated using the Room Impulse Response Generator [23], and T60 values were estimated using Schroeder's backward integration equation. RIRs with microphone and sound source positions less than 0.5 m from room surfaces or closer than 0.2 m to each other were excluded in accordance with the standards [24] [25].

However, several caveats should be considered. First, although 40 RIRs are used at each T60, there is only one room type for each T60. This may not fully reflect real acoustic environments, where room types can vary even for the same T60. Additionally, due to time and computational constraints, we used only 40 RIRs at each T60. Ideally, the number of RIRs should be as large as possible to improve robustness.

Despite these potential limitations, we believe that the reverberation-degraded signals are sufficiently representative of real-world acoustic challenges to provide meaningful insights for our research questions.

3.3 Procedure and Performance Criteria

After applying the test objective intelligibility metrics to the degraded signals and averaging the scores of signals degraded with the same RIR, we obtain 40 intelligibility scores for each T60 value for each test metric. For STIPA, since only one test signal was degraded, the averaging step is omitted, resulting in 40 scores for each T60. In addition to the degraded signals, we also ran STIPA and the test metrics on the clean speech signals and clean test signal to observe how they predict intelligibility in the absence of reverberation.

Levene's Test

To evaluate whether the robustness of test metrics differs between Mandarin and English, we used Levene's test [26] to assess the equality of variances for each test metric across the two languages. Levene's test is robust even when the distributions are not normal, making it suitable for our situation, as the distribution of the predicted scores at each T60 is unknown. The null hypothesis of Levene's test is that the population variances are equal. In our experiments, we used a significance level of 0.05. Therefore, if the resulting p-value is less than 0.05, we reject the null hypothesis.

Kendall's tau coefficient

Kendall's tau coefficient [27], τ , is used to measure the ordinal association between two measured quantities. It requires a smaller sample size to reliably detect correlations compared to other statistical measures, such as Spearman's correlation coefficient [28]. To evaluate how the test metrics perform under low and high reverberation conditions for English, Kendall's tau correlation coefficient is calculated between STIPA and each metric for low T60s, high T60s, and all T60s. We define T60 values of 0.05s, 0.17s, 0.31s, 0.48s, and 0.71s as low reverberation conditions, and 1.17s, 1.92s, 3.15s, and 7.00s as high reverberation conditions. Since each T60 condition includes 40 scores, the sample sizes are 200, 160, and 360 for low reverberation conditions, high reverberation conditions, and the entire range, respectively. Using the method described in [28], detecting correlations stronger than $\tau = 0.6$ with a 95% two-sided confidence interval (CI) and a CI width of 0.2 requires a sample size of approximately 90–100. On the other hand, detecting correlations stronger than $\tau = 0.7$ with a 95% CI and a CI width of 0.1 requires sample size of about 200–250. Therefore, in all three ranges, we can detect correlations stronger than 0.6 with a CI width of 0.2. For the entire range, which has sample size of 360, we can detect correlations stronger than 0.7 with a CI width of 0.1. The null hypothesis of Kendall's tau is that there is no correlation ($\tau = 0$). If the resulting p-value is smaller than the significance level of 0.05, we reject the null hypothesis.

4 Results

In this section we analyze test metric performance under varying reverberation conditions for English and compares scores between Mandarin and English.

4.1 Examining Differences Between Mandarin and English Scores for Test Metrics

Under zero-reverberation conditions, the STIPA score is 0.98, while ESTOI, HASPI, and SIIB$_{\text{Gauss}}$ scores are 1.0, 1.0, and 1335.76, respectively, for both Mandarin and English. Since SIIB$_{\text{Gauss}}$ values above 150 indicate perfect intelligibility, all test metrics predicted perfect intelligibility.

For other reverberation conditions, Figure 1 shows the mean score and standard deviation of the 40 scores at each T60.

For HASPI, it can be observed that its standard deviation becomes significantly larger at 1.17s and 1.92s compared to other T60 values. Overall, the standard deviations of Mandarin and English scores are very close, except at 1.92s, where it is slightly smaller for Mandarin (0.20) compared to English (0.22).

For ESTOI, the standard deviations at each T60 are more consistent, but compared to the mean standard deviation of STIPA (0.02), the standard deviation is still 0.06 larger for both English and Mandarin. The difference between Mandarin and English is minimal, and at the T60 where the standard deviation is largest, the difference is only 0.002.

For SIIB$_{\text{Gauss}}$, the standard deviation is largest at T60 values of 0.05s and 0.17s, decreasing as the T60 value increases. Regarding the differences between languages, at a T60 of 0.05s, English shows a slightly larger standard deviation of 167.22, while for Mandarin, it is 148.38. However, we also need to take into consideration the large magnitude of mean values at that T60.

Table 1 presents the p-values of Levene's test results between Mandarin and English for the test metrics at each T60. We observe that no p-value is smaller than the significance level of 0.05. Therefore, we do not reject the null hypothesis of Levene's test and conclude that there is no difference in variances between Mandarin and English for each test metric.

It is worth noting that while ESTOI and STIPA performed consistently across T60 values, HASPI showed weaker stability at 1.17s and 1.92s, and SIIB$_{\text{Gauss}}$ is less stable at 0.05s and 0.17s. HASPI and ESTOI also had higher average standard deviations compared to STIPA. By combining figure 1 and figure 2, we observe that the means and standard deviations do not differ significantly between Mandarin and English, which aligns with the results of Levene's test.

Table 1: P-Values of Levene's Test Results for Test Metrics Between Mandarin and English at Each T60

	ESTOI	HASPI	SIIB$_{\text{Gauss}}$
0.05s	0.89	0.33	0.68
0.17s	0.86	0.76	0.86
0.31s	0.78	0.43	0.99
0.48s	0.93	0.12	0.90
0.71s	0.95	0.28	0.88
1.17s	0.95	0.90	0.90
0.92s	0.93	0.72	0.95
3.15s	0.97	0.84	0.84
7.0s	0.82	0.72	0.83

4.2 Performance of Test Metrics Under Different Reverberation Conditions for English

Table 2: The Kendalls's tau correlation coefficients between STIPA scores and test metrics scores for English at different T60 ranges

	τ_{Low}	τ_{High}	τ_{All}
ESTOI	0.78	0.68	0.85
HASPI	0.76	0.80	0.87
SIIB$_{\text{Gauss}}$	0.75	0.71	0.84

Table 3: The p-values of Kendalls's tau correlation coefficients between STIPA scores and test metrics scores for English at different T60 ranges

	p-value$_{\text{Low}}$	p-value$_{\text{High}}$	p-value$_{\text{All}}$
ESTOI	1.760e-60	1.222e-37	3.868e-127
HASPI	1.425e-56	3.909e-50	1.171e-134
SIIB$_{\text{Gauss}}$	3.409e-55	3.672e-40	2.673e-123

Table 2 presents the Kendall's tau coefficients between STIPA and test metric scores for English across different ranges, while Table 3 provides the corresponding p-values for τ. All p-values are significantly smaller than the significance level of 0.05, allowing us to reject the null hypothesis and conclude that STIPA scores and test metric scores are correlated.

As mentioned in the previous section, the low range includes T60 values of 0.05s, 0.17s, 0.31s, 0.48s, and 0.71s, while the high range includes 1.17s, 1.92s, 3.15s, and 7.00s. The 'all ranges' category encompasses all these T60 values. In the low reverberation range, performance is similar, with ESTOI achieving the best result, with a τ of 0.78, closely followed by HASPI at 0.76 and SIIB$_{\text{Gauss}}$ at 0.75. In the high reverberation range, HASPI has higher τ than the other metrics, achieving a τ of 0.80, while ESTOI and SIIB$_{\text{Gauss}}$ reached τ values of 0.68 and 0.71, respectively. However, since the confidence interval (CI) width in the high reverberation range is 0.2, it is insufficient to conclude that HASPI outperforms the

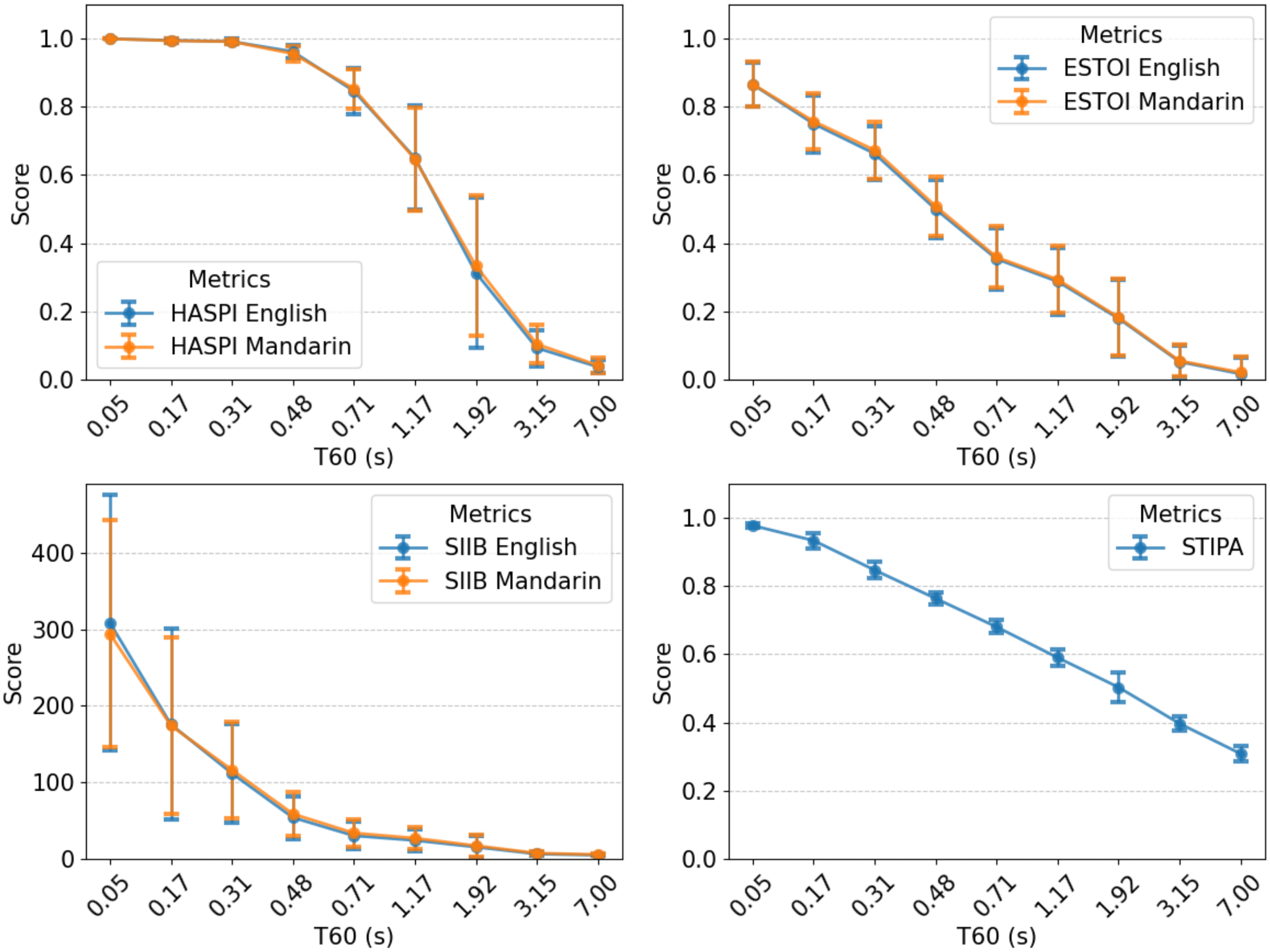

Figure 1: mean scores and standard deviations of test metrics and STIPA. The x-axes are not even step sized

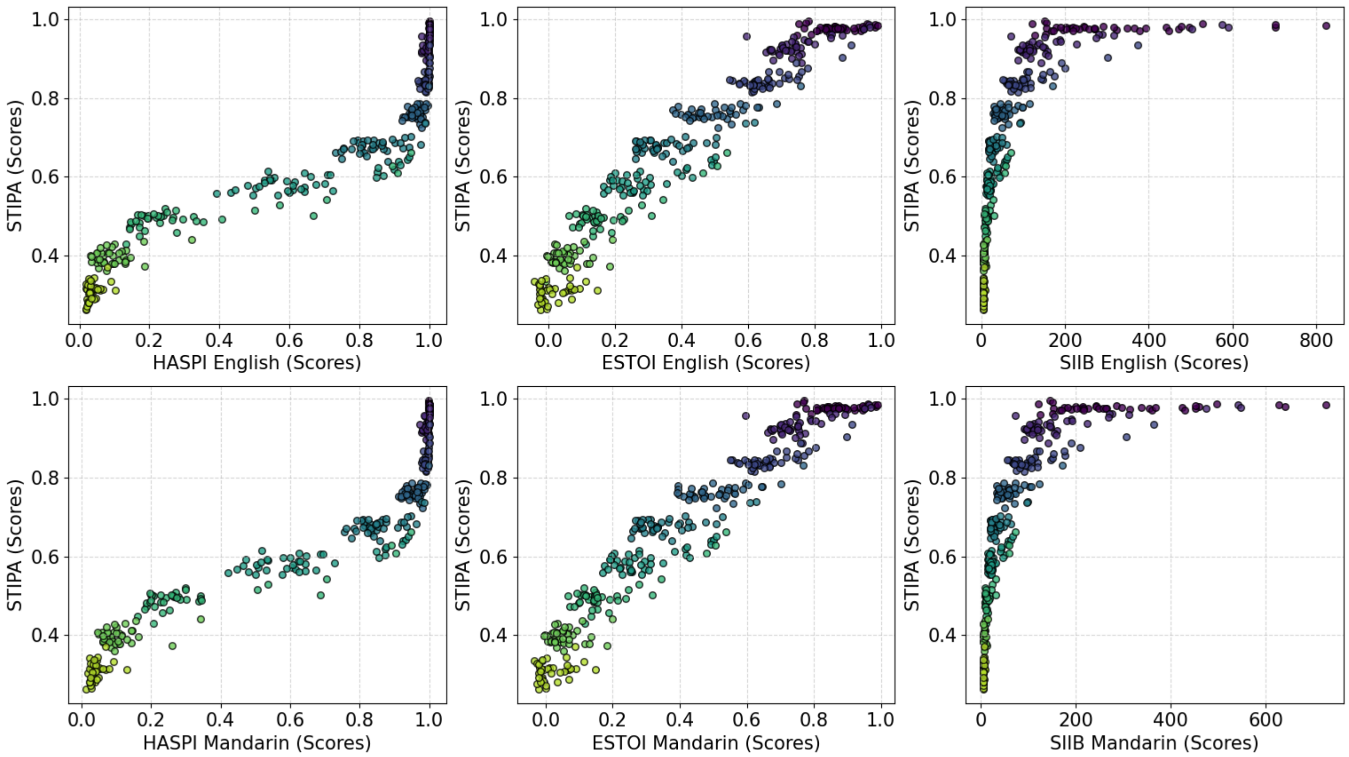

Figure 2: A mapping of test metrics score to STIPA scores for 40 scores at each T60. Scores of same T60 values are of same colors

other two metrics. When all T60 values are included in the τ calculation, HASPI performed best, with a τ of 0.87, followed by ESTOI and $\text{SIIB}_{\text{Gauss}}$. With a CI width of 0.1, due to the higher sample size in this range, we can claim that the performance is similar between the test metrics. These results contradict the theory presented in the study by [15], which found that intrusive metrics perform poorly under high reverberation conditions—except for HASPI—due to temporal blurring caused by severe reverberant distortion, as intrusive metrics rely on temporal alignment. In that study, a T60 of 1.0s was considered strong reverberation. However, as our experimental results demonstrate, this does not hold true for higher T60 values.

4.3 Summary

Summarizing the results, we found that under zero-reverberation conditions, STIPA scores 0.98, while the test metrics indicate perfect intelligibility. Regarding differences in variances at each T60 between English and Mandarin for test metrics, no statistically significant differences were found. In terms of performance under reverberation, all test metrics demonstrate good and similar performance under low reverberation conditions. HASPI appears to outperform the other metrics under high reverberation, but we cannot draw a definitive conclusion due to the CI width of 0.2. When all T60 values are considered, the test metrics all showed strong correlations with STIPA. This finding contradicts the theory proposed in [15].

5 Discussion

Although the test metrics showed no differences in variances between Mandarin and English according to the results of Levene's test, it is also worth noting that there are minimal differences in mean scores at each T60 for the two languages. This finding contradicts the study by [1], which suggests that subjective intelligibility differs under the same room acoustic conditions for different languages, especially in challenging environments. This indicates that there is still room to examine potential biases in the selected test metrics under varying conditions.

By comparing the τ values between STIPA and the test metrics across different T60 ranges, we found that the results do not align with the study by [15], as all test metrics showed similarly strong correlations with STIPA when all T60 values were considered. However, it should be noted that, despite being thoroughly tested for English under reverberant conditions [9], STIPA cannot replace subjective intelligibility tests, which require significantly more time and resources.

Due to time limitations, only 9 T60 values and 40 RIRs per T60 were used to apply degradation, and at each T60 there was only one room type. As a result, we were unable to draw definitive conclusions based on Kendall's tau in the high reverberation range, although differences in τ values were observed. Incorporating a greater variety of RIRs, room types, and additional T60 values would have improved the robustness of the Kendall's tau correlation coefficients and Levene's test results. Furthermore, as mentioned in the previous section, due to dataset limitations, we selected only 84 speakers per dataset. Including more speakers would enhance the robustness of our experimental results.

6 Conclusions and Future Work

Based on our experiment, we found that ESTOI, HASPI, and $\text{SIIB}_{\text{Gauss}}$ show little difference in terms of score variances between Mandarin and English. HASPI, ESTOI, and $\text{SIIB}_{\text{Gauss}}$ also demonstrate similar performance in reverberant conditions (from a T60 of 0.05s to 7s) for English. Due to time and resource limitations, we did not incorporate listening tests in our experiments. However, to obtain more robust results, further research could explore the potential biases of the selected test metrics for different languages and evaluate their performance under reverberant conditions using listening tests.

7 Responsible Research

This research was conducted with a commitment to ethical practices, transparency, and responsibility.

Publicly available datasets, TIMIT and AISHELL-3, were used in this study. These datasets are widely recognized in the speech research community, ensuring their relevance and compliance with licensing agreements. The libraries used in this study are open-source. The implementations of STIPA, ESTOI, HASPI, and $\text{SIIB}_{\text{Gauss}}$ were acquired from publicly available repositories under appropriate licenses. The scripts used to apply OIMs, the resulting .csv files, and the scripts to visualize results can be found on this GitHub page [29]. A README document has been created to explain each file.

The RIRs were generated and applied with the help of supervisors. ChatGPT was used to check for grammar mistakes and to generate templates for running objective metrics and plotting graphs. However, no content in this paper was generated directly using ChatGPT.

References

[1] Laurent Galbrun and Kivanc Kitapci. Speech intelligibility of english, polish, arabic and mandarin under different room acoustic conditions. 114:79–91.

[2] Jont B. Allen. *Articulation and Intelligibility*. Synthesis Lectures on Speech and Audio Processing. Springer International Publishing.

[3] Yong Feng and Fei Chen. Nonintrusive objective measurement of speech intelligibility: A review of methodology. 71:103204.

[4] Tiago H. Falk, Vijay Parsa, Joao F Santos, Kathryn Arehart, Oldooz Hazrati, Rainer Huber, James M. Kates, and Susan Scollie. Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools. 32(2):114–124.

[5] N R French and J C Steinberg. Factors governing the intelligibility of speech sounds. pages 90–119.

[6] Caslav Pavlovic. SII—speech intelligibility index standard: ANSI s3.5 1997. 143(3):1906.

[7] Tammo Houtgast and Herman JM Steeneken. Evaluation of speech transmission channels by using artificial signals. 25(6):355–367. Publisher: European Acoustics Association.

[8] H. J. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. 67(1):318–326.

[9] T. Houtgast, H. Steeneken, and S. V. Wijngaarden. Past, present and future of the speech transmission index. page 92.

[10] Herman Jacobus Steeneken. *On measuring and predicting speech intelligibility*. Steeneken.

[11] Sander J. Van Wijngaarden and Jan A. Verhave. Prediction of speech intelligibility for public address systems in traffic tunnels. 67(4):306–323.

[12] Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. 24(11):2009–2022.

[13] Steven Van Kuyk, W. Bastiaan Kleijn, and Richard C. Hendriks. An instrumental intelligibility metric based on information theory. 25(1):115–119.

[14] James M. Kates and Kathryn H. Arehart. The hearing-aid speech perception index (HASPI) version 2. 131:35–46.

[15] Steven Van Kuyk, W. Bastiaan Kleijn, and Richard Christian Hendriks. An evaluation of intrusive instrumental intelligibility metrics. 26(11):2153–2166.

[16] Lin Yang, Jianping Zhang, and Yonghong Yan. An improved STI method for evaluating mandarin speech intelligibility. In *2008 International Conference on Audio, Language and Image Processing*, pages 102–106. IEEE.

[17] Pavel Záviška, Pavel Rajmic, and Jiří Schimmel. Matlab implementation of STIPA (speech transmission index for public address systems). Publication Title: Journal of the Audio Engineering Society original-date: 2023-11-10T14:14:53Z.

[18] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217. IEEE.

[19] Pariente Manuel. mpariente/pystoi. original-date: 2018-04-18T12:01:22Z.

[20] James M. Kates and Kathryn H. Arehart. The hearing-aid speech perception index (HASPI). 65:75–93.

[21] claritychallenge/clarity: Clarity challenge toolkit - software for building clarity challenge systems.

[22] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. AISHELL-3: A multi-speaker mandarin TTS corpus. In *Interspeech 2021*, pages 2756–2760. ISCA.

[23] Stephen McGovern. Room impulse response generator.

[24] NEN-EN-ISO 3382-2:2008 en.

[25] Maxim de Groot. Estimating reverberation time by a function of intrusive speech intelligibility measures.

[26] Morton B. Brown and Alan B. Forsythe. Robust tests for the equality of variances. 69(346):364–367. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[27] M G Kendall. A NEW MEASURE OF RANK CORRELATION.

[28] Justine O May and Stephen W Looney. Sample size charts for spearman and kendall coefficients. 11.

[29] Jin Mingyi. Myung-kim/CSE3000. https://github.com/Myung-Kim/CSE3000.