An Exploratory Examination of Objective Intelligibility Metrics Under Reverberant Conditions

Author: Mingyi Jin (m.jin-1@student.tudelft.nl)
Supervisors: Jorge Martinez Castaneda, Dimme de Groot

1. Introduction

Speech intelligibility: percentage of words a listener can accurately recognize.
Subjective intelligibility metric: relies on human listeners who evaluate the intelligibility of speech.
Objective intelligibility metric (OIM): uses mathematical models to predict intelligibility.
Intrusive OIM: rely on a clean speech or noise sample as a reference.

- Intrusive intelligibility metrics rely on time alignment between clean and degraded signals, making them overly sensitive to temporal blurring caused by severe reverberant distortion [1]. HASPI (The Hearing-Aid Speech Perception Index) is a exception.
- Significant differences in subjective intelligibility can arise between languages, particularly in acoustically challenging spaces [2].

2. Research Question

Main Question: How do ESTOI, $SIIB_{Gauss}$ and HASPI perform under different reverberant conditions?

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

3. Objective Intelligibility Metrics

Reference Metric:

- **STIPA** (Speech Transmission Index for Public Address Systems): A simpler and faster alternative of STI, designed specifically for testing public address systems. It is reliable for English under reverberant conditions[3].

Test Metrics:

- **ESTOI** (The Extended Short-Time Objective Intelligibility): It is an enhancement of the Short-Time Objective Intelligibility (STOI) algorithm, designed to predict speech intelligibility in environments with highly modulated noise sources or non-linear distortions.
- **$SIIB_{Gauss}$**(Speech Intelligibility in Bits (SIIB) Gaussian): It is computationally faster while maintaining performance levels comparable to SIIB, which estimates speech intelligibility based on information theory by quantifying the amount of information shared between clean and distorted speech signals
- **HASPI** (The Hearing-Aid Speech Perception Index): HASPI 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.

4. Experimental Setup

Datasets: AISHELL-3 for Mandarin, TIMIT for English

Clean signals: 3 utterances x 84 speakers (42 male and 42 female speakers) = 252 utterances each for Mandarin and English

Degraded signals: Each clean signal 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.

Procedure: 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.

Performance Criteria:

- Levene's Test: To evaluate whether the robustness of test metrics differs between Mandarin and English, we used Levene's test to assess the equality of variances for each test metric across the two languages.
- Kendall's tau coefficient: 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.

5. Results

Differences Between Mandarin and English Scores for Test Metrics:

Zero-reverberation: STIPA score is 0.98 (1 means perfect intelligibility), ESTOI, HASPI, and $SIIB_{Gauss}$ predicted perfect intelligibility.

Under reverberation: Levene' Tests results indicate that scores variances are equal between Mandarin and English. Table 1 shows that all resulting p-values are smaller than significance level of 0.05. Slight differences at some T60s can be observed visually from figure 1.

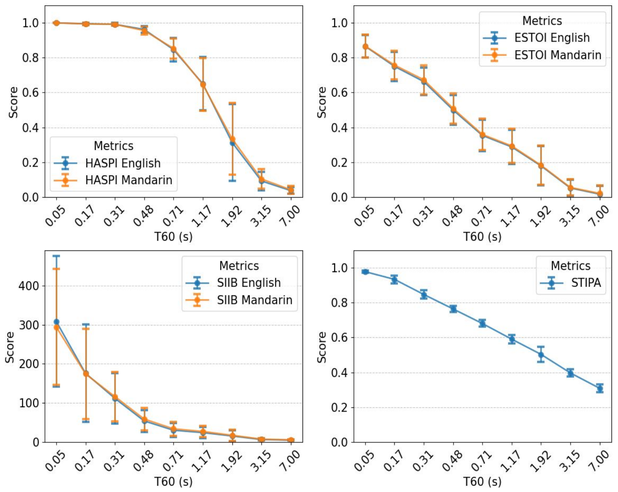

Figure 1: mean scores and standard deviations of test metrics and STIPA. The x-axes are not even step sized

	ESTOI	HASPI	$SIIB_{Gauss}$
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

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

Performance of test metrics for English under reverberation:
From Table 2, it can be observed that In the low reverberation range and entire range, test metrics performances are similar, while in the high reverberation range, HASPI shows better performance than the other metrics. However, since the confidence interval (CI) width in the high reverberation range is 0.2 due to its sample size of 160, it is insufficient to conclude that HASPI outperforms the other two metrics. In low reverberation range and entire range CI width is 0.1 because of sample sizes of 200 and 360.

	τ_{Low}	τ_{High}	τ_{All}
ESTOI	0.78	0.68	0.85
HASPI	0.76	0.80	0.87
$SIIB_{Gauss}$	0.75	0.71	0.84

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

As shown in Table 3, the p-values are small enough to demonstrate strong correlation between test metrics and STIPA for English.

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

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

6. Limitations

- Despite being thoroughly tested for English under reverberant conditions [3], STIPA cannot replace subjective intelligibility tests.
- Only 9 T60 values and 40 RIRs per T60 were used to apply degradation, and at each T60 there was only one room type. Also due to dataset limitations, we selected only 84 speakers per dataset.
- Including more speakers and incorporating a greater variety of RIRs, room types, and additional T60 values would have improved reliability of the experimental results.

7. Conclusions and Future Work

ESTOI, HASPI, and $SIIB_{Gauss}$ show little difference in terms of score variances between Mandarin and English.
HASPI, ESTOI, and $SIIB_{Gauss}$ also demonstrate similar performance in reverberant conditions (from a T60 of 0.05s to 7s) for English.
Further research could be done by using listening tests.

References

[1] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Evaluation of Intrusive Instrumental Intelligibility Metrics," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 26, no. 11, pp. 2153–2166, Nov. 2018, doi: 10.1109/TASLP.2018.2856374.
[2] L. Galbrun and K. Kitapci, "Speech intelligibility of English, Polish, Arabic and Mandarin under different room acoustic conditions," Applied Acoustics, vol. 114, pp. 79–91, Dec. 2016, doi: 10.1016/j.apacoust.2016.07.003.
[3] T. Houtgast, H. Steeneken, and S. V. Wijngaarden, "Past, present and future of the speech transmission index," 2002. Accessed: Jan. 15, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Past%2C-present-and-future-of-the-speech-transmission-Houtgast-Steeneken/ca7e863f8ac584cb0631a703745c96e76b7c1f4f