Author: Maxim de Groot
21-5-2024

Estimating Reverberation Time by a Function of Intrusive Speech Intelligibility Measures

Introduction

Estimating reverberation time (T60) is beneficial considering that Automatic Speech Recognition systems can adapt to reverberation using estimated T60 values.

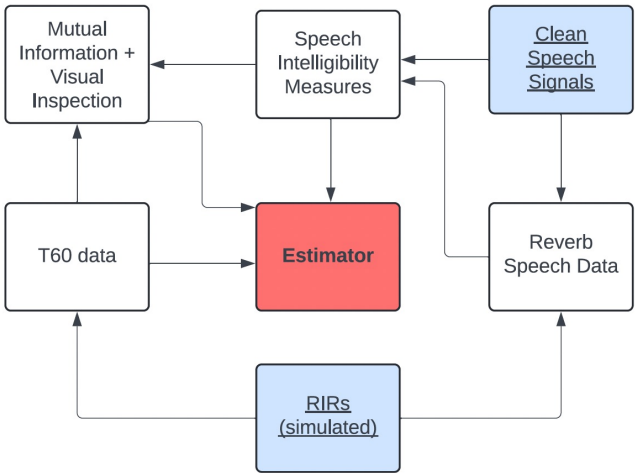
- Currently T60 estimation is mostly done using machine learning models
- A statistical estimation approach exploiting the relation of the T60 and speech intelligibility can be more efficient (less data and faster).

Speech Intelligibility Measures

Instead of formal listening tests you can use objective speech intelligibility measures (SIMs) to obtain an intelligibility score. Some of these measures' performance are known to be affected by reverberation. Using the SIIB, SIIB^Gauss, STOI and ESTOI SIMs we answer the question:

"Can intrusive SIMs be used to estimate the T60 in speech signals by using a statistical estimator?"

Estimator Construction



- Simulated RIRs (Image Source Method)
- Mutual information (KNN estimation)
- Find formula with best curve fit by minimizing MSE using Broyden–Fletcher–Goldfarb–Shanno algorithm
- Evaluate using MSE and MAE on simulated and real dataset, also with additive noise

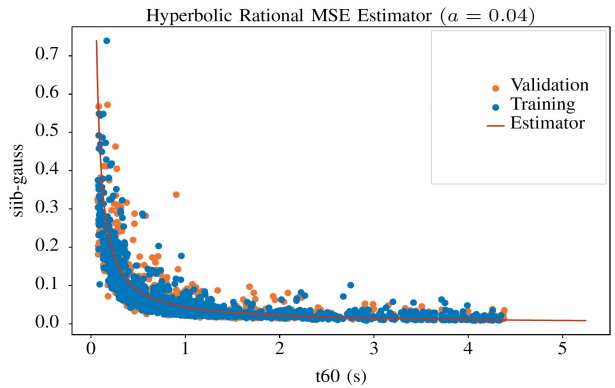
Results

Mutual Information between each SIM and the T60 for training data.

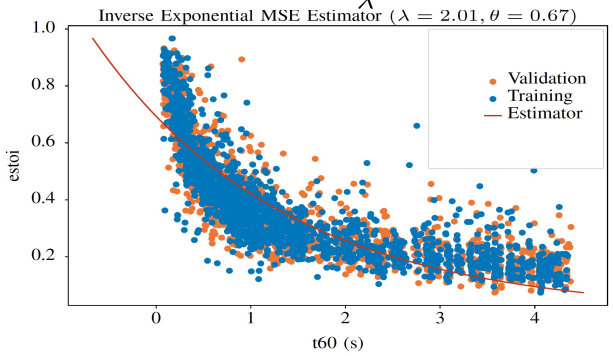
SIIB	SIIB^{Gauss}	STOI	ESTOI
0.962	1.042	0.681	0.968

Resulting estimators of minimizing MSE for SIIB^Gauss (SIIB is similar) and ESTOI (STOI is similar):

$$f(x) = \frac{a}{x}$$



$$g(x) = \theta + \lambda(\log(\frac{1}{\lambda}) - \log(x))$$



Evaluation of the estimator on simulated and real RIRs (2000 simulated, 300 real samples)

	SIIB	SIIB^{Gauss}	STOI	ESTOI
MSE Simulated	0.417	0.353	0.552	0.396
MAE Simulated	0.474	0.403	0.563	0.479
MSE Real	0.336	0.323	0.099	0.677
MAE Real	0.358	0.352	0.240	0.341

Evaluation with additive Gaussian noise with specified signal-to-noise ratios (100 samples)

	SIIB	SIIB^{Gauss}	STOI	ESTOI
MSE (SNR = 10 dB)	3.644	3.05	0.795	1.50
MAE (SNR = 10 dB)	1.469	1.385	0.761	0.852
MSE (SNR = 20 dB)	2.970	3.27	0.751	0.481
MAE (SNR = 20 dB)	1.330	1.399	0.695	0.552
MSE (SNR = 30 dB)	2.760	3.768	0.742	1.206
MAE (SNR = 30 dB)	1.327	1.507	0.660	0.817

Conclusions and Future Work

SOTA achieves accuracies ~ 0.03 MSE

- Performance not comparable but seems good on real RIRs
- Not very noise robust → add preprocessing
- Combining multiple SIMs into single estimator (SIIB^Gauss and ESTOI)
- Using different way of finding best fit curve (e.g. maximum likelihood)
- Use non-intrusive measures (only requires reverberant speech)
- Evaluate bias for gender, dialects etc.

References

Scan QR code for reference list

Responsible Professor: Jorge Martinez Castaneda
Daily Supervisor: Dimme de Groot