SPEECH METRICS IN REVERBERANT ENVIRONMENTS

Hugh Hopper

1 INTRODUCTION

Performance of telecommunications systems can be related to intelligibility and subjective quality. Both of these are related to human perception and therefore must be measured with listening tests. However, these tests are costly and time consuming, not to mention tedious for the listeners. Objective metrics, such as noise levels or distortion, have clear links to the subjective performance and are easily measured, but do not always provide a true picture of the system performance. Therefore, many objective metrics which estimate subjective performance have been developed.

Although most telecoms metrics consider near field speech signals, such as from handset or headset microphones, far field speech is becoming more common in conferencing systems and calling capabilities of smart home devices. In the far field, the effect of room reverberation is significant. This paper investigates the behaviour of the objective speech metrics on reverberant speech signals. This should demonstrate whether and under what conditions the metrics may be used.

2 BACKGROUND

In order to understand whether the metrics have a behaviour that represents subjective perception by human listeners, actual subjective data must be considered. However, actually gathering listening test data is beyond the scope of this study. The results presented by Cueille et al.[1] show the effect of noise and reverberation on speech intelligibility. From that work, the results considered here will be those from normal hearing listeners where reverberation is applied only to the speech signal, not the noise signal.

The results from this study are given in rationalized arcsine units (RAU) which represent a scaled and fitted psychometric function relating to speech intelligibility. The effect of varying the signal-to-noise ratio (SNR) on the RAU intelligibility for various reverberation times is shown in Figure 1. From these results, the following characteristics can be inferred:

1. Speech intelligibility decreases with signal to noise ratio, i.e., with increasing noise level.
2. The relation between SNR and intelligibility follows a sigmoid-type curve.
3. At equal SNR, signals with longer reverberation time have worse intelligibility.
4. The effect of reverberation is more pronounced at higher SNR.

For the purpose of this work, it will be assumed that to represent human perception of speech, a performance metric should have these characteristics.

Furthermore, some approximate quantitative results can be extracted:

1. The transition from maximum to minimum intelligibility covers an SNR range of 10 to 20 dB.
2. The mid-point of the value range (approx 60 RAU in this case) occurs at an SNR of -7 to -3 dB.
3. A high intelligibility score (100 RAU) is achieved at an SNR of 5 dB and -2 dB for reverberation times of 1.5 s and 0.15 s respectively: the increase in RT is equivalent to a 7 dB drop in SNR.
4. A low intelligibility score (25 RAU) is achieved at an SNR of -9 dB and -11 dB for reverberation times of 1.5 s and 0.15 s respectively: the increase in RT is equivalent to a 2 dB drop in SNR.

These results are unlikely to be generalizable so should not be used a definitive test of the correctness of the metric results. However, it can be of interest to see if the results are similar.

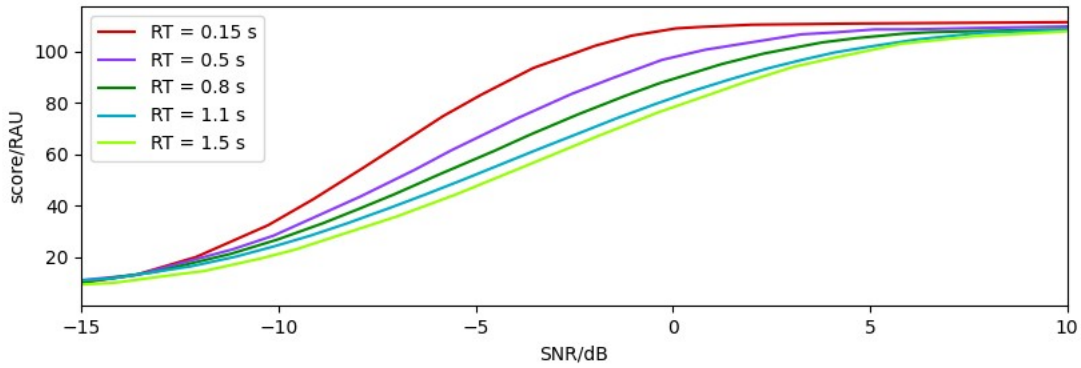

Figure 1: Increasing SNR and reducing RT causes a higher speech intelligibility. The RAU values are based on scaled and fitted psychometric functions. Results after Cueille et al.[1]

3 METHODOLOGY

3.1 METRICS

Several objective measures of speech quality are available. For the sake of the present study, a few have been selected: PESQ, STOI and DNSMOS. These are freely available and have existing Python implementations. Additionally, they represent a variety of the metrics available due to their differing approaches.

Perceptual evaluation of speech quality[2] (PESQ) was developed as an ITU standard P.862. It requires a copy of the clean speech as a reference. The signals are analysed using psychoacoustic loudness in the Bark spectral domain and the perceptual model accounts for masking and other hearing features. PESQ has been superseded by POLQA, but that requires a proprietary license so it has not been considered here.

Short-term objective intelligibility[3] (STOI) also uses frequency domain analysis between a clean reference signal and the degraded signal. In this case correlation is used to measure the similarity between the two signals. The temporal envelopes are compared in each frequency band and then a time-frequency weighting is used to calculate the final value.

Deep Noise Suppression Mean Opinion Score[4] (DNSMOS) was developed to support the Deep Noise Suppression Challenge which has featured at InterSpeech and ICASSP conferences. It uses a deep learning method to estimate the mean opinion score directly from the degraded speech only. This means DNSMOS could be applied where the clean signal is not available.

Note that STOI aims to measure intelligibility and the other two estimate MOS which relates to perceived quality. The estimated MOS values vary from 1 to 5 whereas STOI outputs a value from 0 to 1. Therefore, for this study the STOI value will be scaled to match the MOS range which means that the response of these metrics to noise and reverberation can be analysed and compared.

3.2 SIMULATION

To test the performance of the metrics, a wide range of speech signals has been generated. The clean speech is taken from the LibriSpeech dataset[5]. The corpus is based on recordings of audio books so the speech signals have low noise and reverberation. Noise signals have been taken from the QUT-NOISE dataset[6] which features a range of noise signals from domestic environments, inside vehicles and in public spaces. Finally, room impulse responses were taken from the MIT IR Survey[7,8] which contains measured impulse responses from a wide range of real spaces.

To generate the test audio, the selected speech signal is convolved with the impulse response. The level of the resulting signal is measured using the active speech level (ASL) based on ITU P.56 which calculates speech level without the silent gaps between syllables and words[9]. The level of the noise signal is calculated based on the A-weighted Leq. Finally, the reverberation time of each impulse response is the T20 value from gradient of a linear regression on the inverse integration of the impulse[10].

The speech level is adjusted so that it equals -24 dBov, to avoid any influence of overall amplitude on the speech quality. The noise level is then set to create the desired SNR. The output signal is created based on the sum of the scaled reverberant speech and scaled noise. For PESQ and STOI, the unaltered speech signal without noise or reverberation is presented as the reference signal.

For negative SNR values, with high noise level, the amplitude of the final signal may be reduced in order to avoid clipping. Although this will reduce the speech level, it is assumed this is preferable to distortion on the noise signal, especially as SNR is the key variable of interest.

The simulations have been conducted with 5 different speech signals and 5 different noise signals in all combinations. A total of 20 different impulse responses covering a range of reverb times from 0.1 s to 1.8 s was used, plus an anechoic case where no impulse is applied. An SNR range of -15 to 30 dB was evaluated in 5 dB steps and then from 40 to 90 dB in 10 dB steps.

Each of the metrics is calculated based on the resulting output signal. The results are grouped into 3 bands of reverberation time: low is between 0.1 and 0.5 s, medium for between 0.5 and 1.0 s and high for between 1.0 and 2.0 s. The results are also grouped by SNR and then the arithmetic mean is calculated for the three metrics.

4 RESULTS AND DISCUSSION

The evaluation of the metrics against signals with reverberant speech and noise, shown in Figure 2, allows a comparison between metrics and also an assessment against the criteria discussed in section 2. For all three metrics, the basic qualitative criteria are fulfilled: at lower SNR the MOS is decreased, the trend between SNR and MOS follows a sigmoid-type curve and at equal SNR the higher reverberation time causes lower MOS. Additionally, the effect of reverberation is more pronounced at higher SNR, which manifests as greater horizontal distance between the three curves at the higher end of the SNR range.

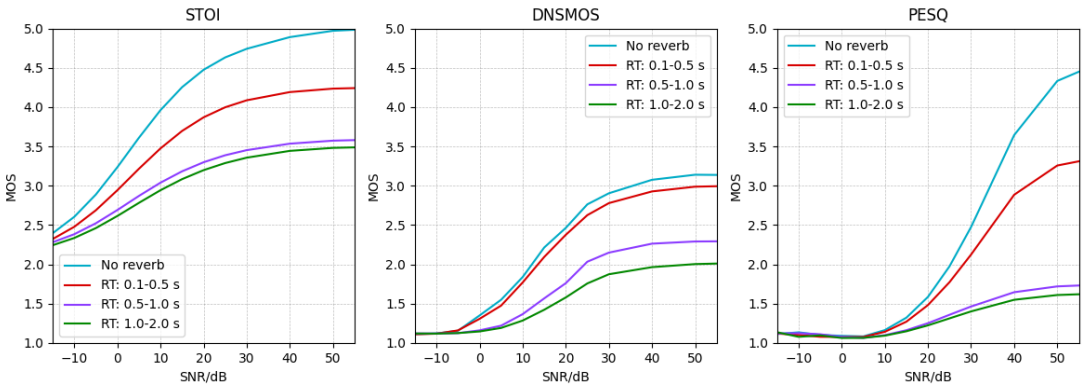

Figure 2: The effect on the estimated MOS for each metric for a range of reverberation times and signal to noise levels.

Although the overall qualitative performance of the metrics seems to behave as expected, the three metrics clearly perform very differently. To understand these differences, the quantitative measures listed in section 2 can also be analysed.

Firstly, the transition region from maximum to minimum MOS covered approximatley 10-20 dB in the data in section 2, with that range of variation depending on the reverberation time. However, the results from these metrics show a much wider transition region of 40 dB being fairly similar between the three metrics. Although this may not be the most accurate representation of human perception, it could be considered a useful feature as a wider transition region will allow a greater range of signals to be meaningfully compared.

Secondly, the midpoint of the transition region in Figure 1 was between -7 and -3 dB. For these metrics the midpoint is significantly higher. STOI has a midpoint between 0 and 10 dB, DNSMOS between 10 and 20 dB and PESQ at between 20 and 30 dB. It is worth recalling that these metrics are not aiming to produce identical results, so the different locations of the transition regions are expected. However, the differences are quite large and should be considered in relation to the signals being analysed when choosing an appropriate metric.

Finally, the quantitative effect of reverberation can be considered. The data from section 2 showed that an increase in reverberation time from 0.15 s to 1.5 s caused a loss in intelligibility equivalent to a change in SNR of maximum 7 dB. This occurs in cases where the SNR is high, and the effect of reverberation is decreased when the SNR is lower.

The results for the objective metrics show much greater sensitivity to reverberation. The equivalent SNR change for the increase in reverberation time is at least 30 dB for all three metrics. In particular for PESQ, even at SNR greater than 40 dB the application of short reverberation times causes a drastic reduction in the estimated MOS.

An additional point worth noting is that DNSMOS produced a maximum value of 3 across all results in this study. This is due to the inherent quality of the LibriSpeech dataset and the particular utterances that were chosen out of it. This is because DNSMOS does not use a reference signal so it estimates the absolute quality of the signal.

5 CONCLUSIONS AND FUTURE WORK

The aim of this study was to investigate the behaviour of objective speech metrics—specifically, PESQ, STOI, and DNSMOS—under the influence of reverberant speech in various signal-to-noise ratio (SNR) conditions. The research aimed to determine whether these metrics can accurately approximate human perception of speech quality in reverberant environments.

The results demonstrated that the selected metrics fulfil basic qualitative criteria aligning with human perception, i.e., decreasing SNR and increasing reverberation results in lower estimated quality. Further analysis of the metrics revealed distinct differences in their quantitative performance. The three metrics had different sensitivities to noise, with a different critical SNR marking the transition from high to low quality estimate. STOI had the lowest at around 0-10 dB, then DNSMOS at 10-20 dB and PESQ the highest at 20-30 dB. These ranges may be used to decide which metric applies to given signals of interest.

Additionally, the metrics are more sensitive to reverberation than would be expected from human listeners. In particular, increasing the reverberation time from 0.15 s to 1.5 s resulted in a decrease in estimated quality that would correspond to a reduction in SNR of 30 dB, where listening test data would imply less than 10 dB. Furthermore, PESQ was very sensitive to any reverberation, even with RT60 or less than 0.5 s the estimated quality was drastically reduced.

Given these results, some general recommendations can be given. PESQ should only be used for near field signals as even low reverberation times caused significant drops in estimated quality.

Both STOI and DNSMOS provide results which cover a useful range of scenarios, with DNSMOS covering a higher SNR range. The results are consistent enough that within the same environment the metrics could be used to compare signals. However, the effect of reverberation on these metrics means they cannot be used to compare signals between reverberant environments or when considering the performance of de-reveberation algorithms.

The results are perhaps unsurprising given that these metrics were likely developed for near-field signals. However, given the greater presence of far-field voice communications for teleconferencing and from smart home devices, it is important to quantify these results to motivate the development of metrics specifically designed for far-field cases.

For future work, it would be interesting to consider other metrics such as POLQA and VISQOL. It would be useful to have more detailed listening test data in a range of reverberant environments, especially when considering subjective quality. Additionally, a greater range of noise and speech input signals could be considered, in particular higher quality speech data could improve the range of results. Finally, a longer term goal would be the development of far-field specific performance metrics.

6 REFERENCES

1. Cueille, Raphael, Mathieu Lavandier, and Nicolas Grimault. "Effects of reverberation on speech intelligibility in noise for hearing-impaired listeners." Royal Society Open Science 9.8 (2022).
2. Rix, Antony W., et al. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs." ICASSP, Proc of. (2001).
3. Taal, Cees H., et al. "An algorithm for intelligibility prediction of time–frequency weighted noisy speech." IEEE Transactions on Audio, Speech, and Language Processing 19.7 (2011).
4. Reddy, Chandan KA, Vishak Gopal, and Ross Cutler. "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors." ICASSP, Proc. of. (2021).
5. Panayotov, Vassil, et al. "Librispeech: an ASR corpus based on public domain audio books." ICASSP, Proc of. (2015).
6. Dean, David, et al. "The QUT-NOISE-TIMIT corpus for evaluation of voice activity detection algorithms." Conference of the International Speech Communication Association, (2010).
7. Traer, James, and Josh H. McDermott. "Statistics of natural reverberation enable perceptual separation of sound and space." Proceedings of the National Academy of Sciences 113.48 (2016).
8. Traer and McDermott. "IR Survey." URL: mcdermottlab.mit.edu/Reverb/IR_Survey.html. Accessed 17 Oct. 2023.
9. "Objective measurement of active speech level". ITU P.56, (2011).
10. Schroeder, Manfred R. "Integrated-impulse method measuring sound decay without using impulses." The Journal of the Acoustical Society of America 66.2 (1979).