Audio Engineering Society

Convention e-Brief 248

Presented at the 140th Convention
2016 June 4–7, Paris, France

This Engineering Brief was selected on the basis of a submitted synopsis. The author is solely responsible for its presentation, and the AES takes no responsibility for the contents. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Audio Engineering Society.

The influence of discrete arriving reflections on perceived intelligibility and STI measurements

Ross Hammond[1], Peter Mapp[2], and Adam Hill[1]

[1]*Department of Engineering, University of Derby*
[2]*Peter Mapp Associates*

Correspondence should be addressed to Ross Hammond (`rosshammond@mail.com`)

ABSTRACT

The most widely used objective intelligibility measurement method, the Speech Transmission Index (STI), does not completely match the highly complex auditory perception and human hearing system. Investigations were made into the impact of discrete reflections (with varying arrival times and amplitudes) on STI scores, subjective intelligibility and the subjective ‘annoyance factor’. This allows the effect of comb filtering on the modulation transfer function matrix to be displayed, as well as demonstrates how the perceptual effects of a discrete delay cause subjective ‘annoyance’, that is not necessarily mirrored by STI. This work provides evidence showing why STI should not be the sole verification method within public address and emergency announcement systems, where temporal properties also need thoughtful consideration.

1 Introduction

It is essential for public address systems, especially those used for emergency announcement, to have an acceptable level of intelligibility. An accurate method of measurement is therefore significantly important, as an overestimation could pose serious safety concerns. The Speech Transmission Index (STI) is one of the most widely accepted methods in the electroacoustic design and installation industry, confirmed by its inclusion in the current standard for fire and evacuation procedures [1], and is often used as the sole verification technique for proof of an adequately performing sound system. STI is effective for many types of distortion which degrade intelligibility, but as it does not completely match the highly complex human hearing system flaws arise in its mechanism.

The standardised method [2] allows for a single value rating, between 0 and 1, to predict the intelligibility of a sound transmission system by measuring the extent of preserved fluctuations in speech. The input signals form a collection of sinusoidally varying intensities at modulation frequencies from 0.63Hz to 12.5Hz, at one third octave intervals. The 14 modulations are captured across 125Hz to 8KHz at octave centre frequencies, forming a 7 by 14 matrix of modulation indices. Each measurement captures the reduction in a received signal’s modulation depth, compared with the original transmitted signal. Reductions occur due to the influence of reverberation, reflections or background noise. The array of modulation transfer functions are averaged to form the Transmission Index (TI) for each frequency band. The TI values are applied with an intelligibility weighting function and combined to produce a single STI value. Alternatively, the ‘indirect’ method involves computing the STI from an impulse response [3].

A high level discrete reflection will produce a comb filtering effect observed in the frequency and modulation frequency domain, which can critically affect the modulation transfer function and STI [4], where a short delay time will correspond to a high modulation frequency and vice versa. If the delayed signal and the direct sound are synchronised to cause an interference null at a modulation frequency, an erroneous result can be produced.

Many psychoacoustic phenomena related to discrete arriving reflections exist, such as the fusion of early reflections with the direct sound aiding intelligibility [5], or the delay time and level relationship of a reflection having a direct correlation with its perception and disturbance [6]. Both could produce significant differences between the interpretation of a sound and the corresponding STI.

Previous research [7] has shown that in the absence of other variables, a single discrete delay up to 500ms will not reduce intelligibility beyond a certain measurable threshold. However, the subjective impression of speech quality reduces at a different rate, suggesting an innate difference. The degree to which speech can be easily, or comfortably, understood is especially important in the case of lecture halls, religious buildings and places where spoken word is a primary function (although maybe less so for emergency announcement systems), as it will determine long term concentration levels and listener comfort.

2 Methods

Two types of tests were conducted to show differences between objective measurements and subjective impression. STI measurements were made within an anechoic chamber at the University of Birmingham, with an additional artificial delayed signal at different times between 0ms and 500ms. Indirect measurements were made via maximum length sequence (MLS), using Clio 10 software/hardware package by Audiomatica [8], and extracted impulse response files were analysed with EASERA [9]. A Behringer X32 digital mixing console [10], eliminating analogue component degradation, distributed the MLS signal to two full range active loudspeakers, delaying the secondary signal as necessary. Both loudspeakers were placed at a two meter distance to the measurement microphone with the 'direct signal' loudspeaker on axis and the 'delayed signal' loudspeaker positioned 30 degrees off axis anticlockwise. Both loudspeakers were set to a reference level of 65dBA, with the delayed loudspeaker played at 0dB, -3dB, -6dB and -10dB periodically for each delay time. The theoretical MTF matrix was also obtained for a collection of delay times via simulations in Matlab.

Subject based headphone tests were conducted to examine the subjective effects of a discrete artificial delay. A mean opinion score (MOS) test was modified to incorporate 'perceived intelligibility', as well as the usual 'speech quality' and 'listener effort', which are each based on a five point scale. 1 represents 'completely unintelligible', 'bad quality/very annoying' and 'no meaning understood with feasible effort' respectively and 5 represents 'completely intelligible', 'excellent quality/imperceptible' and 'no effort required' respectively. Pre-recorded speech material was used from an audiobook read by Stephen Fry [11], and a collection from librivox.org [12] which consisted of 4, 5.5 and 7 syllables per second speech rates. Audio files were divided into approximately 15 second (+/- 1s) phrases, without adjustments to time or pitch. Conditions were chosen so that no more than a total of 2 seconds existed in silent intervals between speech content in each phrase. The rate of speech for each phrase was chosen to be within +/- 0.5 syllables per second and +/- 2 words per minute, to keep a consistent delivery rate. Any silences greater than or equal to 150ms were omitted from calculations. Each 15 second phrase was presented only once throughout the entire testing process, and randomly assigned to a condition for a single listening participant, eliminating periodicities, recognisable transients and repeated phrases from influencing results. All tests took place within a controlled environment at the University of Derby, with noise levels not exceeding 30dBA. Participants consisted of 20 young adults with no known hearing impairments. Signals were processed in Logic Pro X [13], with use of the 'delay designer' plugin to create the desired delay times, utilising binaural processing to create a 0 degrees direct sound at 2 meters and 30 degrees delayed sound at 2 meters with no elevation.

3 Results

Expected comb filtering was clear from the frequency response of the MLS measurements, which is also apparent in the modulation domain and within the Transmission Index findings (Fig 1), due to the synchronisation of delay time and modulation frequencies.

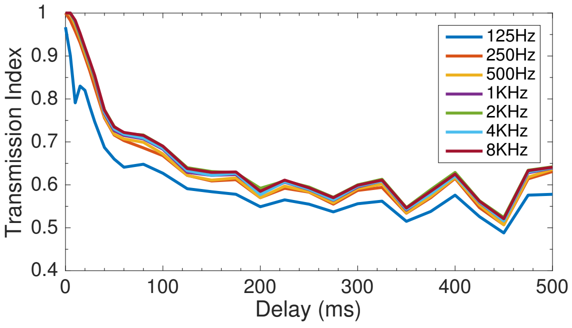

Figure 1 - Transmission Index values for each frequency band for 0ms to 500ms

An example shown in Fig 2, represents the modulation reduction for a 100ms delay, which has a null at 5Hz (200ms). The mid point of this modulation, at maximum amplitude, will be delayed by 100ms to the end of the modulation, resulting in no modulation depth. The comb pattern continues as delay time is increased and the delay begins to synchronise with lower modulation frequencies.

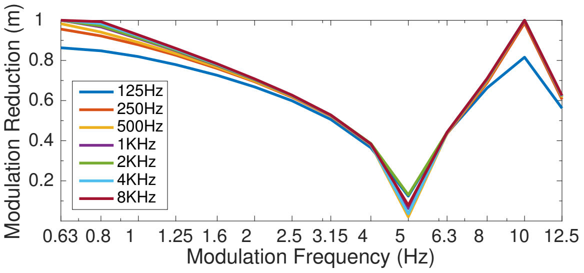

Figure 2 - Modulation reduction for 14 modulation frequencies, for each frequency band at a 100ms delay time

Since STI only takes 14 modulation frequencies (or samples) into account, as delay time and comb filter rate increases, a greater variance exists between adjacent delay times. This is seen in Fig 3 which shows 225 and 250ms delay times which have a very similar trend in the high modulation frequency sample rate. The modulation reduction between the two delay times vary in the actual modulation frequencies measured. This is a significant value at 10Hz for example, which contributes to the reduced STI score at 250ms. A higher comb rate (and delay time) will increase the chance of errors introduced into the TI value.

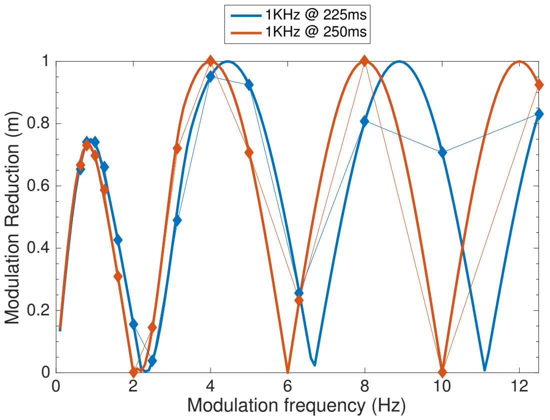

Figure 3 - A comparison of the theoretical modulation reduction of 225 and 250ms, for both high sample rate (smooth curve) and actual sample rate (straight lines) at 1KHz sound frequency band

An aliasing effect is possible, as seen in Fig 4, where a 500ms delay time shows an extreme example with comb filter peaks which synchronise with the sampled modulation frequencies creating a high TI value. This is supported by the STI results shown in Fig 5, where a significant peak is seen at 500ms.

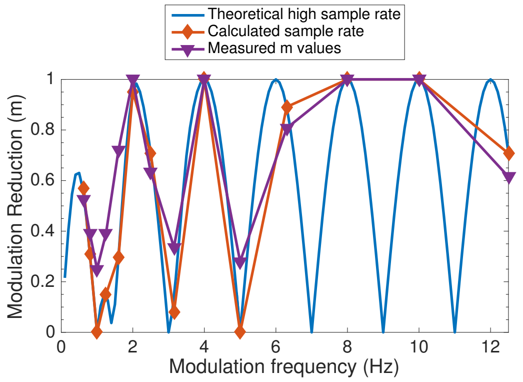

Figure 4 - Modulation reduction for 500ms, including theoretical high sample rate, calculated sample rate and values derived from measurements

Final STI values (Fig 5) follow similar trends to the TI values/modulation transfer function matrix, since all frequency bands follow the same result as delays were an exact duplicate of the signal, so the intelligibility weighting function has little impact. As expected, reduced delay levels increase overall STI values and reduce the impact of the comb filtering effect. Measurements do not solely represent the masking caused from long delay times, but are influenced by the synchronisation between delay time and modulation frequency.

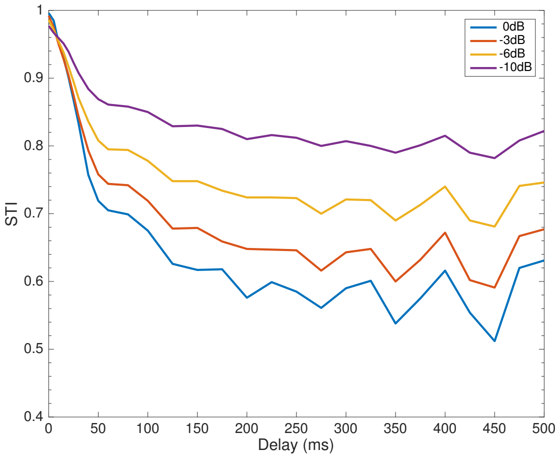

Figure 5 - STI scores when a discrete delay is applied at varying times between 0ms and 500ms at 4 levels

The perceived intelligibility from subjective MOS results (Fig 6) follows a different trend to STI scores, showing a consistent level. However, it can be clearly observed that there is an innate difference between intelligibility and quality, suggesting the 'annoyance' of an echo should be treated as a separate factor. A discrete reflection, in the absence of other distortions, may not reduce intelligibility by a significant amount, but the effort required to understand the speech decreases as delay time is incremented. The comb filter effect seen in STI scores is not represented in the subjective interpretation.

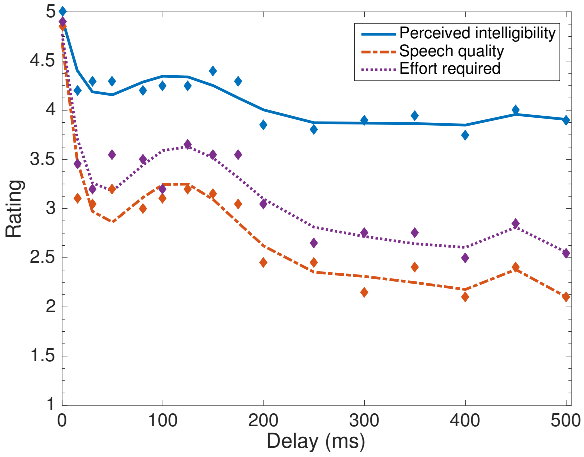

Figure 6 - Mean opinion score results for a 4sps talker rate for delay times between 0 and 500ms, at 0dB delay level

Speech rate MOS results (Fig 7) demonstrate how an increase in speech rate does not simply equate to a linear decrease in perceived quality across all delay times, contradictory to work from [6]. A correlation exists between synchronisation of syllables per second and delay time with reduced subjective impression, creating non-linear trends.

As an example, the 7sps rate would on average introduce a new syllable every 143ms which corresponds with the null produced at 125-150ms. Results mostly conform to the findings of previous research that a slower rate of speech is more intelligible. However, at certain delay times, a faster rate of speech is perceived to have better intelligibility and quality, suggesting that a faster rate of speech may be beneficial in certain troublesome situations, such as 5.5 sps having a significantly higher quality than 4 sps at 250ms.

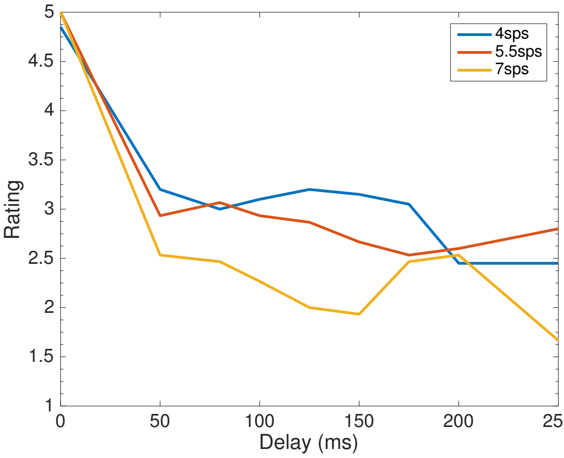

Figure 7 - Mean opinion score 'quality' results for 4, 5.5 and 7sps talker rates for delay times between 0 and 250ms, at 0dB delay level

4 Discussion

For a discrete reflection with long delay time, STI measurements will differ according to the synchronicity of the modulation frequencies and delay time, suggesting an unreliable measurement method within a space with this type of distortion. The rate of the comb within the MTF increases as delay time is incremented, creating greater variation between delay intervals, which also produces a greater chance of aliasing in the 'sampled' modulation frequencies. The STI and MTF findings provide evidence of this effect, as comb filtering shows most significant alterations in higher delay times.

Despite the ability to understand everything that is said by a talker, it will become uncomfortable to listen in environments with strong echoes for long listening periods, as conditions with ‘bad quality/large amount of effort required’ ratings are also rated as ‘moderately to completely intelligible’. As actual intelligibility [7], perceived intelligibility and speech quality do not follow the trends of STI scores, an issue is apparent with STI measurements in this type of distortion. This is due to the mechanistic issue with the STI method, where STI scores are determined purely by the synchronisation of delay time and modulation frequencies, as well as the potential psychoacoustic differences.

MOS trends mostly follow work from [6], [14], who show a reflection of equal level will be perceived at 30ms and have a 50 percent listener disturbance rate at 60ms, respectively. A steep decrease is present in the subjective MOS results at 30ms, but increases for a peak at 125ms before linearly declining. When the speech rate is taken into account at 4sps (on average, each syllable is spoken every 250ms), a syncopated pattern between syllables and delay may be influencing the results and reinforcing the speech around 125ms. The steady decline beyond this point is otherwise supported by previous research. Evidence supports the importance of the rate of speech, which is also not considered by measured STI scores. Within emergency announcements and pre-recorded messages, results suggest that in the presence of high level discrete delays, an effort should be made to avoid a rate of speech which will combine with it negatively.

5 Summary

STI should not be used as a sole verification method in spaces with high level discrete reflections, due to the differences found between measurements and subjective impression in the absence of other distortion. Temporal properties should also be considered using an echogram, energy time curve or even observing the MTF for further validation. The perceived quality of speech is also dependent on the synchronisation between speech rate and delay time, where results suggest a faster speech rate can improve the perceived intelligibility and impression for certain troublesome combinations. Experimentation into a modified method which incorporates a greater amount of modulation frequencies is currently in progress, to try to overcome the mechanistic issues found. Further work investigating the effects of multiple arriving reflections, different echo densities and spatial information will be insightful for the development of an improved method.

References

[1] ISO, “7240-24:2010: Fire detection and fire alarm systems.” 2010.

[2] IEC, “60268-16: Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index,” 2011.

[3] Schroeder, M., “Modulation transfer functions: Definition and measurement,” *Acta Acustica united with Acustica*, 49(3), pp. 179–182, 1981.

[4] Mapp, P., “Modifying STI to Better Reflect Subjective Impression,” in *Audio Engineering Society Conference: 21st International Conference: Architectural Acoustics and Sound Reinforcement*, 2002.

[5] Bradley, J., Sato, H., and Picard, M., “On the importance of early reflections for speech in rooms,” *The Journal of the Acoustical Society of America*, 113(6), pp. 3233 – 3244, 2003.

[6] Haas, H., “The Influence of a Single Echo on the Audibility of Speech,” *J. Audio Eng. Soc*, 20(2), pp. 146–159, 1972.

[7] Hammond, R., “The influence of discrete arriving reflections on perceived intelligibility and STI measurements,” Undergraduate Dissertation, University of Derby, 2016.

[8] Audiomatica, “http://www.audiomatica.com,” 2016.

[9] AFMG, “http://easera.afmg.eu,” 2016.

[10] Music-Group, “www.music-group.com/p/P0ASF,” 2016.

[11] Apple, “https://itunes.apple.com/gb/audiobook/fry-chronicles-autobiography/id392834710,” 2010.

[12] Librivox, “https://librivox.org,” 2016.

[13] Apple, “http://www.apple.com/uk/logic-pro/,” 2016.

[14] Meyer, E. and Schodder, G. R., “Göttinger Nachrichten.” *Math. Phys.*, Kl(11a), 1962.