Audio Engineering Society

Convention Paper 9629

Presented at the 141st Convention
2016 September 29–October 2 Los Angeles, USA

The influence of discrete arriving reflections on perceived intelligibility and Speech Transmission Index 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 potential 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 over the range 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 number of simplified STI measurements have also been developed, including STIPA which incorporates 2 modulation frequencies per octave band simultaneously. Fewer measurements, although practically beneficial, compromise accuracy and introduce potential errors [4].

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 [5], 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 [6], or the delay time and level relationship of a reflection having a direct correlation with its perception and disturbance [7]. Both could produce significant differences between the interpretation of a sound and the corresponding STI.

As well as intelligibility, 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

To find the effects of discrete arriving reflections on STI measurements, intelligibility and subjective impression a number of testing methods were used allowing any differences to be identified. STI measurements were made within an anechoic chamber at the University of Birmingham and an auditorium at the University of Derby, with an additional artificial delayed signal at different times between 0ms and 500ms, found in Table 1. 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 anti-clockwise. 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.

0ms	40ms	175ms	350ms
5ms	50ms	200ms	375ms
10ms	60ms	225ms	400ms
15ms	80ms	250ms	425ms
20ms	100ms	275ms	450ms
25ms	125ms	300ms	475ms
30ms	150ms	325ms	500ms

Table 1. The collection of tested delay times

The theoretical modulation transfer function (MTF) matrix and frequency response was also obtained for a collection of delay times via simulations in Matlab. As well as validating the conducted measurements, this allowed a comparison between a high sample rate and actual sample rate in the MTF domain.

A modified rhyme test (MRT) allowed the intelligibility of the same conditions to be tested within the same auditorium, with an artificial delay added using the same methodology as the STI measurements. Although less sensitive than other approaches, an MRT allowed for a greater amount of conditions to be tested. The speech material was pre-recorded in a vocal booth and read by a female with neutral London-British accent. Facilities used include Avid pro-tools digital audio workstation [11], Focusrite 828 audio interface [12] and AKG C414 microphone [13], with closed back headphones used for monitoring, and recorded at 48kHz sample rate and 24 bit depth. Recording each test word and carrier sentence separately three times, and merging the test words within the carrier sentence, allowed for a consistent delivery rate of 4 syllables per second (sps), as well as the option to audition test words. The sentences were separated by a three second interval. Listening tests were conducted with 10 young adults, in groups of two or three, all of whom were unaware of possessing any hearing impairments. 4 fifty-word MRT tests were conducted with each group. Artificial delays were played back at 0dB, in reference to the direct signal, with delay times found in Table 2.

Delay	Group 1	Group 2	Group 3	Group 4
0ms	✓	✓	✓	✓
15ms		✓		
30ms			✓	
50ms				✓
80ms	✓			
100ms		✓		
150ms			✓	
200ms				✓
250ms	✓			
300ms		✓		
400ms			✓	
500ms	✓			✓

Table 2. Delay times used during MRTs

Subject based tests employing headphones 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 [14], and a collection from librivox.org [15] 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 [16], 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 Modulation Transfer Function Matrix

Expected comb filtering was clear from the frequency response of the MLS measurements, which is also apparent in the modulation frequency domain and within the Transmission Index findings (Fig 1), due to the synchronisation of delay time and modulation frequencies. The 125Hz band, although following the same overall trend, is lower than the other frequency bands due to noise. This is confirmed by the reduced 125Hz values throughout all modulation frequencies for individual modulation transfer functions.

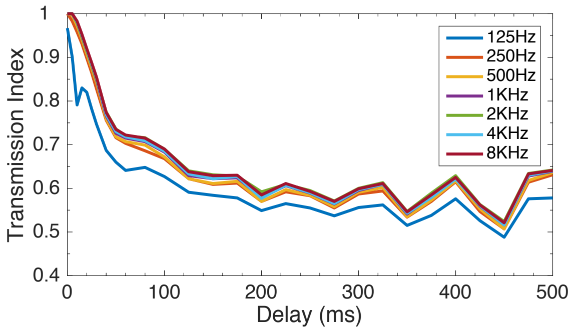

Figure 1. Transmission Index values for each frequency band for 0ms to 500ms

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.

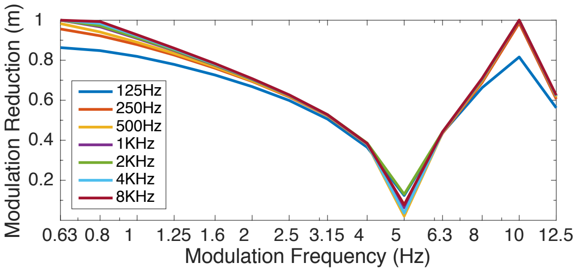

Figure 2. Modulation reduction for 14 modulation frequencies, for each frequency band at a 100ms delay time

As delay time increases, the null seen in modulation reduction becomes less exaggerated, until synchronisation starts to occur with the next modulation frequency (4Hz). The jagged curve seen in TI values (Fig 1) is a direct result of this, as the delay alternates between in and out of syncronisation. The comb pattern continues as delay time is increased, until a second null appears as the delay time synchronises with two modulation frequencies.

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.

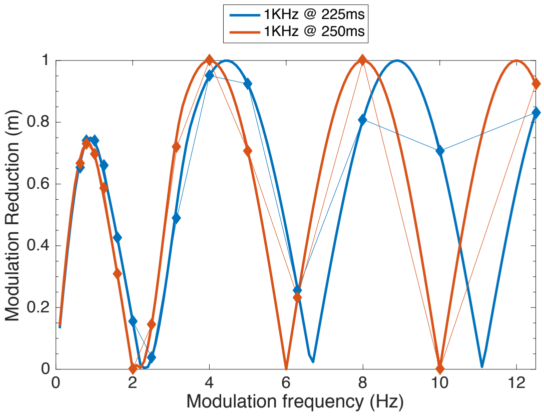

Figure 3. A comparison of the theoretical modulation reduction of 225 and 250ms, for both high sample rate (smooth curve) and actual sample rate (straight lines) at 1kHz sound frequency band

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. In context, this increase could raise the STI to an apparently acceptable level, highlighting the extent of this inaccuracy.

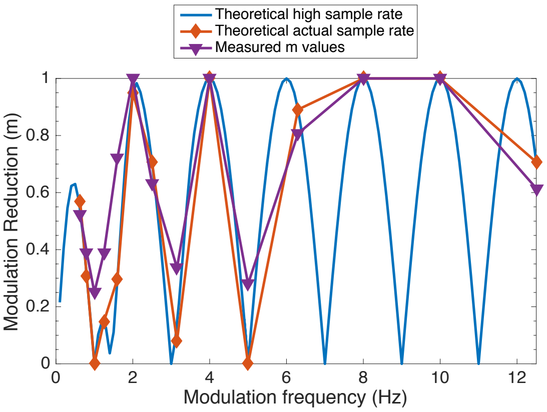

Figure 4. Modulation reduction for 500ms, including theoretical high sample rate, calculated sample rate and values derived from measurements

4 Speech Transmission Index Results

Final STI values follow similar trends to the TI values, 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. The effects of delay level on STI results can be seen in Fig 5. As expected, reduced delay levels increase overall STI values and reduce the impact of the comb filtering effect. It can be observed that for delay times between 0ms to ~20ms, a higher delay level increases the STI score. This corresponds to the perceptual impact of early reflections increasing the perceived signal to noise ratio and intelligibility. As delay time begins to increase from 0ms, only the highest modulation frequencies exhibit a reduction whilst the delay reinforces the lower modulation frequencies which overall creates an increase in STI.

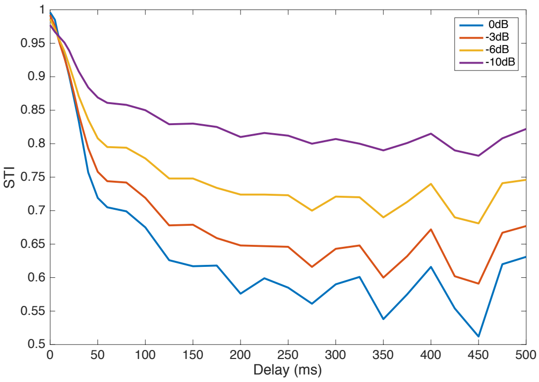

Figure 5. STI results taken within an anechoic chamber for four delay levels

The effects of delay angle and measurement microphone distance can be seen in Fig 6 and 7, for measurements within the auditoria. Trends display similar comb filtering effects for both 30 and 60 degrees, as the monaural measurement would exhibit a similar summation. Comb filtering becomes less exaggerated for a further distance although this is due to an increase in external reflections from the room. Fig 6 (0dB) shows the STI scores which will correspond to the MRT tests.

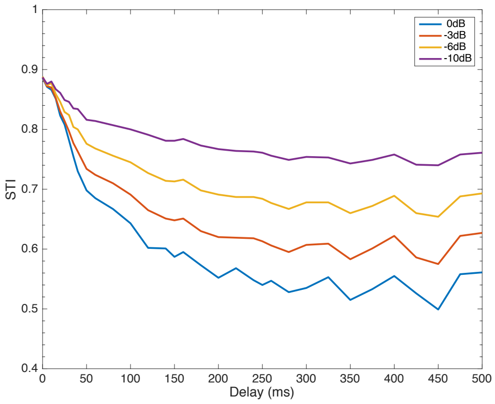

Figure 6. STI results taken within an auditorium for four delay levels, with a 60 degree delay angle and 2m microphone distance

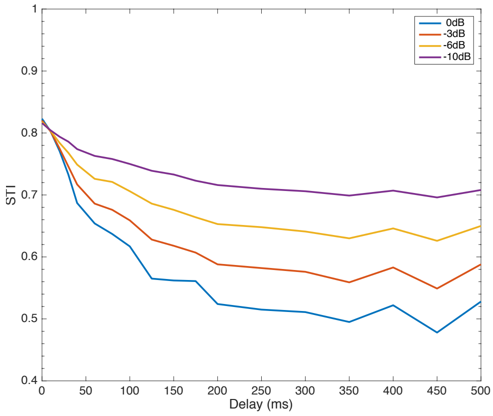

Figure 7. STI results taken within an auditorium for four delay levels, with a 60 degree delay angle and 4m microphone distance

5 STIPA

Evidence has shown a mechanistic issue with using STI measurements within this type of distortion (high level, discrete reflections beyond ~100ms). However, STIPA is a measurement method which could posses further errors. As STIPA measurements only take two modulation frequencies into account for each frequency band, it suggests that results could vary significantly from the corresponding STI result. As Transmission Index (TI) values are weighted according to an intelligibility weighting curve, and the TI values form an average of the modulation reduction values, it would suggest that at a point where a null is present in the MTF, it would reduce the corresponding frequency band by a significant amount compared with STI measurements. This can be seen in Fig 8 which shows the theoretical differences between modulation frequency samples for each frequency band for a 250ms delay.

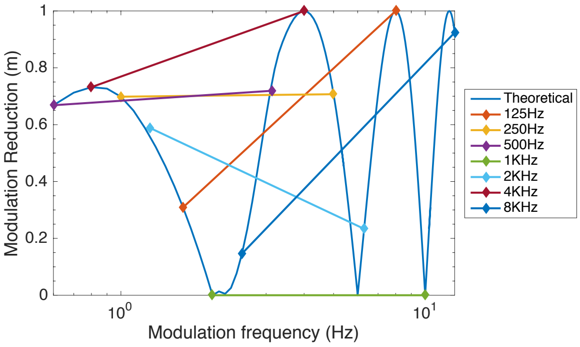

Figure 8. Modulation frequencies for STIPA with a 250ms delay time

The corresponding TI values are seen in Fig 9, which shows significant differences when compared with the TI values for an STI measurement. This is especially apparent for 1kHz which reduces by 0.55.

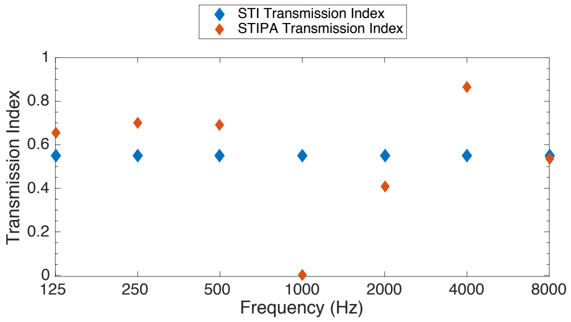

Figure 9. A comparison of TI values for STI and STIPA for a 250ms delay time

The balance of TI values for STIPA will average to form the TI value for STI. However, since the intelligibility weighting function means the contribution from each frequency is different, the distribution of TI values determines the differences, showing how STIPA is fundamentally flawed beyond STI measurements and should not be used in environments with high level, discrete arriving reflections with long delay times.

6 Discrete Reflections and Intelligibility

MRT results, shown in Fig 10, show a consistently high intelligibility level, despite a very strong perceived echo. Corresponding STI scores which drop from ~0.85 to ~0.55 STI would suggest there should be a greater reduction in intelligibility, which demonstrates a difference between objective and subjective testing within this type of distortion. MRT testing is not particularly sensitive to high intelligibility levels, but the high percentages and consistent scores do suggest a difference. The high scores create a fundamental difficulty in determining the effects of delay time on actual intelligibility, since such a small deviation is shown throughout.

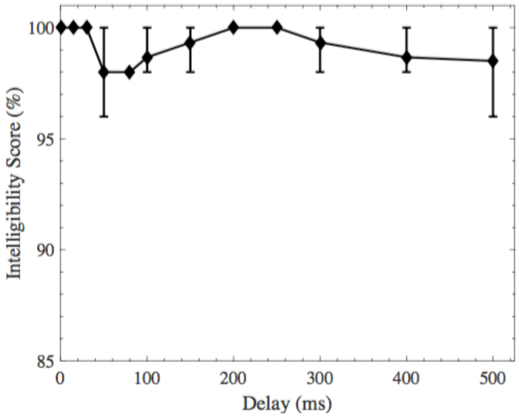

Figure 10. Modified rhyme test intelligibility scores

7 Subjective Impression

The perceived intelligibility from headphone based subjective MOS results (Fig 11) follow 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 increases as delay time is incremented. The comb filter effect seen in STI scores is not represented in the subjective interpretation.

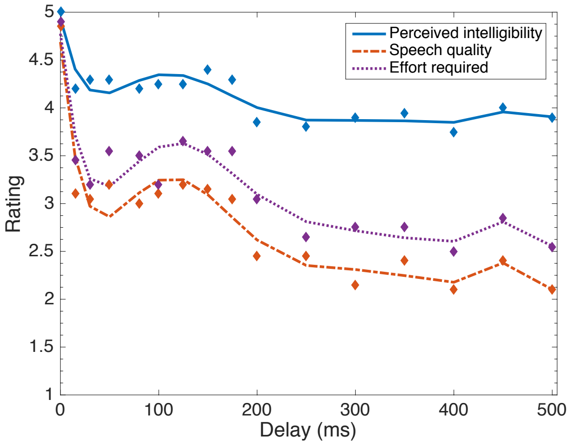

Figure 11. Mean opinion score results for a 4sps talker rate for delay times between 0 and 500ms, at 0dB delay level

Fig 12 shows the effect of delay level on quality. At 0dB, it is observed from past research [7], [17], that an echo will be perceived at 30ms and has a 50% listener disturbance rate at 60ms. A steep decrease is present in the 0dB MOS results at ~30ms, but increases for a peak at 125ms before linearly declining, which does not line up with previous research. However, when the speech rate is taken into account at 4sps (showing on average each syllable is spoken every 250ms), a syncopated pattern between syllable and delay may be influencing the results and reinforcing the speech at certain delay times. The steady decline beyond this point supports previous research.

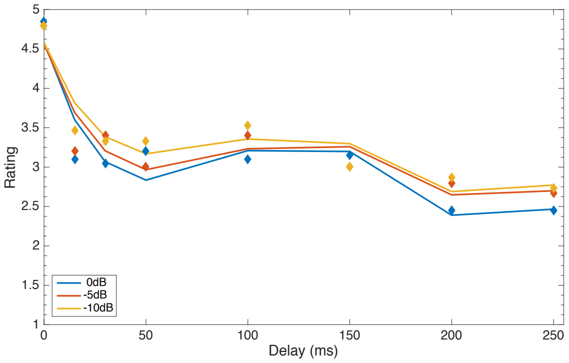

Figure 12. Mean opinion score results for a 4sps talker rate for delay times between 0 and 250ms, at 3 delay levels

8 Rate of Speech

Speech rate MOS results (Fig 13 and 14) demonstrate how an increase in speech rate does not simply equate to a linear decrease in perceived intelligibility and quality across all delay times, contradictory to work from [7]. A correlation is seen to exist between the synchronisation of syllables per second and delay time with reduced subjective impression, creating these non-linear trends.

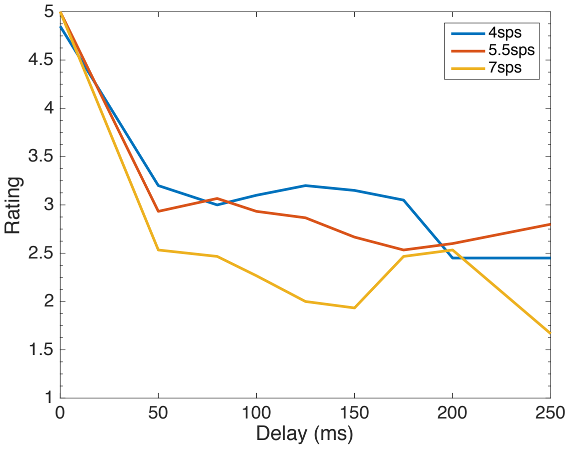

Figure 13. Mean opinion score 'quality' results for 4, 5.5 and 7sps talker rates

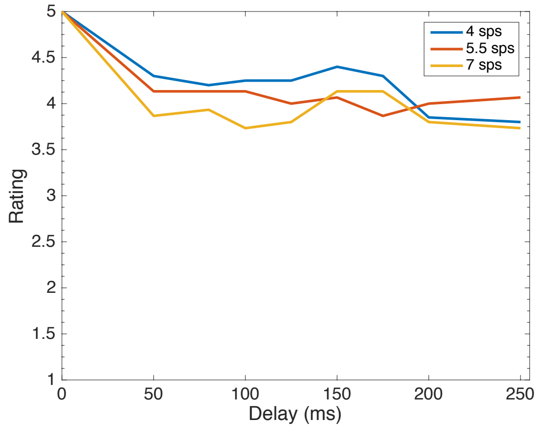

Figure 14. Mean opinion score ‘perceived intelligibility’ results for 4, 5.5 and 7sps talker rates

As an example, the 7sps rate would on average introduce a new syllable every 143ms which corresponds with the null produced at 125-150ms. Subjective impression improves as delay time increases beyond this point. 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 5.5sps having a significantly higher quality than 4sps at 250ms.

9 Summary

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, 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.

Evidence also supports the importance of the rate of speech, which is not considered by the measured STI scores. In 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. A faster speech rate has been shown to improve the perceived intelligibility and impression for certain troublesome combinations. As only three talkers and speech rates were analysed, and as speech signals can be divided into elements beyond syllables, further investigations into the combination of discrete reflections and speech phonetics will be useful for continued advice for pre-recorded message content and delivery rate when in the presence of high level, discrete reflections.

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.

Experimentation into a modified method which incorporates a greater amount of modulation frequencies or bypasses the late arriving reflection 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., "Is STIPA a Robust Measure of Speech Intelligibility Performance?" In: *Audio Engineering Society Convention 118,* 2005.

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

[6] 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.*

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

[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] Avid, "http://www.avid.com/en/pro-tools", 2016.

[12] Focusrite, https://us.focusrite.com/mic-pres/isa-828#", 2016.

[13] AKG, "http://cloud.akg.com/7744/c414xls_xlii_manual.pdf", 2016.

[14] Apple,"https://itunes.apple.com/gb/audiobook/fry-chronicles-autobiography/id392834710," 2010.

[15] Librivox, "https://librivox.org," 2016.

[16] Apple, "http://www.apple.com/uk/logic-pro/," 2016.

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